CEDRE

Cycle des Évaluations Disciplinaires Réalisées sur Échantillons

Rapport technique

Histoire-géographie et enseignement moral et civique 2017

Collège

Auteurs :

Stéphane BERTON
Marion LE CAM
Louis-Marie NINNIN
Thierry ROCHER
Ronan VOURC'H

Bureau de l'évaluation des élèves
DEPP - Direction de l'évaluation, de la prospective et de la performance
Ministère de l'éducation nationale et de la jeunesse

Juin 2019

Table des matières

Introduction

La Direction de l'Evaluation, de la Prospective et de la Performance (DEPP) met en place des dispositifs d'évaluation des acquis des élèves reposant sur des épreuves standardisées. Elle est également maître d'oeuvre pour la France des évaluations internationales telles que PIRLS ou PISA. Ces programmes d'évaluations sont des outils d'observation des acquis des élèves pour le pilotage d'ensemble du système éducatif (Trosseille & Rocher, 2015). Les évaluations du CEDRE (Cycle d'Evaluations Disciplinaires Réalisées sur Echantillons) révèlent ainsi, en référence aux programmes scolaires, les objectifs atteints et ceux qui ne le sont pas. Ces évaluations doivent permettre d'agir au niveau national sur les programmes des disciplines, sur l'organisation des apprentissages, sur les contextes de l'enseignement, sur des populations caractérisées.

Leur méthodologie de construction s'appuie sur les méthodes de la mesure en éducation et sur des modélisations psychométriques. Ces évaluations concernent de larges échantillons représentatifs d'établissements, de classes et d'élèves. Elles permettent d'établir des comparaisons temporelles afin de suivre l'évolution des performances du système éducatif.

Ce rapport présente l'ensemble des méthodes qui sont employées pour réaliser les évaluations du cycle CEDRE, en balayant des aspects aussi divers que la construction des épreuves, la sélection des échantillons ou bien la modélisation des résultats. L'objectif est de rendre accessible les fondements méthodologiques de ces évaluations, dans un souci de transparence. La publication de ce rapport fait d'ailleurs partie des engagements pris par la DEPP dans le cadre du processus de certification des évaluations du cycle CEDRE.

1 Cadre d'évaluation

1.1 Objectifs

Le cycle des évaluations disciplinaires réalisées sur échantillon (CEDRE) établit des bilans nationaux des acquis des élèves en fin d'école et en fin de collège. Il couvre les compétences des élèves dans la plupart des domaines disciplinaires au regard des objectifs fixés par les programmes officiels. La présentation des résultats permet de situer les performances des élèves sur des échelles de niveau allant de la maîtrise pratiquement complète de ces compétences à une maîtrise bien moins assurée, voire très faible, de celles-ci. Renouvelées régulièrement, ces évaluations permettent de répondre à la question de l'évolution du niveau des élèves au fil du temps.

Ces évaluations n'ont pas valeur de délivrance de diplômes, ni d'examen de passage ou d'attestation de niveau ; elles donnent une photographie instantanée de ce que savent et savent faire les élèves à la fin d'un cursus scolaire. En ce sens, il s'agit bien d'un bilan. Destinées à être renouvelées périodiquement, ces évaluations-bilans permettent également de disposer d'un suivi de l'évolution des acquis des élèves dans le temps. Pour cette raison, les épreuves ne peuvent pas être rendues publiques car, devant être en grande partie reprises lors des cycles d'évaluation suivants, elles ne doivent pas servir d'exercices dans les classes.

Ces évaluations apportent un éclairage qui intéresse tous les niveaux du système éducatif, des décideurs aux enseignants sur le terrain, en passant par les formateurs d'enseignants : elles informent sur les compétences et les connaissances des élèves à la fin d'un cursus, elles éclairent sur l'attitude et la représentation des élèves à l'égard de la discipline ; elles interrogent les pratiques d'enseignement au regard des programmes ; elles contribuent à enrichir la réflexion générale sur l'efficacité et la performance de notre système éducatif. Ces évaluations étant proposées à des échantillons statistiquement représentatifs de la population scolaire de France métropolitaine, aucun résultat par élève ne peut être calculé.

CEDRE a été initié en 2003 avec l'évaluation des compétences générales. Afin d'assurer une comparabilité dans le temps, l'évaluation est reprise pour chaque discipline selon un cycle de six ans jusqu'en 2012 et de cinq ans depuis 2012 (tableau 1).

1.2 Connaissances et compétences visées

1.2.1 Les programmes de référence

L'évaluation CEDRE en fin de collège en histoire, géographie et enseignement moral et civique a pour objectif de faire le point des connaissances et des com-

Tableau 1 – Cycle des évaluations disciplinaires réalisées sur échantillons (CEDRE) depuis 2003

Discipline évaluée	Début du cycle	Reprises	
Maîtrise de la langue et compétences générales	2003	2009	2015
Langues étrangères	2004	2010	2016
Attitude à l'égard de la vie en société	2005	–	–
Histoire, géographie et éducation civique	2006	2012	2017
Sciences	2007	2013	2018
Mathématiques	2008	2014	2019

pétences des élèves tant sur le plan des savoirs que des savoir-faire d'une part, et de mesurer l'évolution de ces connaissances et compétences entre 2006, 2012 et 2017 d'autre part.

Les connaissances et compétences telles qu'elles sont définies dans les programmes officiels constituent le cadre de cette évaluation.

En mai 2006 comme en mai 2012, les élèves de troisième ont été évalués à partir des programmes de 1995 (classes de sixième, cinquième et quatrième) et de 1998 (classes de troisième).

En mai 2017 en revanche, les élèves de troisième ont été interrogés en référence aux programmes de 2008 pour leurs classes de la sixième à la quatrième incluse, et aux programmes de 2015 entrés en vigueur à la rentrée 2016 pour leur classe de troisième.

1.2.2 Tableau de compétences

En 2017, comme en 2012 et en 2006, une même grille de compétences est utilisée pour les trois champs de cette évaluation : histoire, géographie et enseignement moral et civique. La plupart des intitulés sont restés identiques entre les trois points de mesure, ce qui permet d'assurer la comparaison entre les prises d'informations.

Cette grille reste en effet pertinente pour faire place aux renouvellements liés aux programmes de 2015 et aux nouvelles compétences qu'ils redéfinissent et qui, pour rappel, ont concerné les élèves évalués en 2017 pour leur année de troisième seulement. Par exemple, les compétences " se repérer dans le temps " et " se repérer dans l'espace " des programmes de 2015 renvoient à la compétence CEDRE " identifier-localiser "; " raisonner, justifier une démarche et les choix effectués " se superpose à " interpréter "; " analyser et comprendre un document " correspond, selon le cas, à " identifier-sélectionner l'information

Tableau 2 – Tableau des compétences

	Histoire	**Géographie**	**Enseignement moral et civique**
Identifier - Localiser	HL	GL	EL
Identifier - Décrire	HD	GD	ED
Identifier - Sélectionner	HS	GS	ES
Traiter l'information - Classer	HC	GC	EC
Traiter l'information - Mettre en relation	HM	GM	EM
Interpréter - Généraliser	HG	GG	EG
Interpréter - Argumenter	HA	GA	EA
Interpréter - Réaliser	HR	GR	ER

Note de lecture : Chaque compétence est définie par une lettre (L pour localiser, D pour décrire,..) chaque matière est définie de façon identique (H pour histoire, G pour géographie et E pour enseignement moral et civique). Le croisement de la matière et de la compétence est matérialisé par un couple de lettres, associé à l'item produit.

" ou " traiter l'information ", voire " interpréter ". Enfin, " pratiquer différents langages en histoire-géographie " renvoie à " interpréter-réaliser ". Seules deux nouvelles compétences de 2015, à savoir " s'informer dans le monde du numérique " et " coopérer et mutualiser ", n'ont pas pu être évaluées dans le cadre de CEDRE 2017. Il faut cependant préciser que ces compétences de 2015 sont destinées à prendre la place de la grille des compétences CEDRE pour la prochaine échéance de 2022, du fait notamment des nouvelles opportunités d'une évaluation qui sera intégralement opérée sur support numérique, et sous réserve que les programmes de 2015, ou du moins les compétences qu'ils définissent, restent en vigueur jusqu'à ce quatrième temps de mesure.

Par ailleurs, pour mieux prendre en compte la dimension déclarative constitutive de la notion même de compétence, chacun des items du corpus de l'évaluation de 2017 a aussi été indexé selon une seconde grille (Cf. Tableau 3). Elle définit les connaissances évaluées selon une ou plusieurs de ces composantes : concepts ou notions, autre vocabulaire, repères (historiques et géographiques), acteurs ou actrices (historiques ou spatiaux).

Tableau 3 – Grille des connaissances associées aux compétences (pour deux exemples d'items)

items	Concepts et notions	Autre vocabulaire	Repères	Acteurs ou actrices
C3HGG2102	mondialisation	métropole		
C3HHD850101	Monarchie (absolue)			Louis XIV/le Roi-Soleil

L'évaluation CEDRE se décline en deux volets : le premier concerne des items présentés dans un cahier d'évaluation ; le second correspond à celui passé sur ordinateur.

— 1er volet : évaluation " papier " : Il s'agit pour l'élève de répondre à des

unités présentées sur un support " papier ". Une unité se compose d'un ou de plusieurs documents que l'élève doit utiliser pour répondre aux questions. Les unités sont regroupées dans treize blocs ; les treize blocs dans treize cahiers. C'est l'évaluation de référence pour le descriptif des acquis des élèves ainsi que l'instrument permettant la comparaison entre les différents points de mesure.

— 2nd volet : évaluation " numérique " : Il s'agit pour les élèves d'un second échantillon, distinct du précédent, de répondre à des unités présentées sur un support " électronique ". L'unité propose le plus souvent un document à l'élève et des questions afférentes à ce dernier. Le document peut prendre la forme d'une image ou d'un texte, d'un ensemble de textes et d'images regroupés dans une liseuse, d'une vidéo ou d'un croquis dont la légende est à compléter (en glisser-déposer). Les unités sont " ventilées " dans six blocs ; les six blocs dans quinze modules. Il faut noter que cette partie n'est pas prise en compte pour la comparaison temporelle. Mais elle fait l'objet d'une analyse spécifique. Après calcul des scores et publication de la note d'information, ces items numériques ont été projetés sur l'échelle papier, afin d'enrichir, en vue de publications complémentaires, la description des niveaux de maîtrise des compétences évaluées pour chacun des six groupes distingués selon leur score global au test parmi les élèves interrogés.

1.2.3 Calendrier de l'évaluation

La passation de l'évaluation CEDRE histoire, géographie et enseignement moral et civique en fin de collège a eu lieu en mai 2017. La méthodologie mise en oeuvre par la DEPP en la matière, de la conception de l'évaluation à la passation des épreuves puis à l'analyse et à la publication des résultats, s'étale sur trois années selon le schéma suivant :

Tableau 4 – Les étapes de la réalisation de l'évaluation

1ère année	2ème année	3ème année
2015-2016	2016-2017	2017-2018
Étape 1	Étape 2	Étape 3
Expérimentation	Évaluation	Analyse et publication des résultats
Passation mai 2016	Passation mai 2017	
Objectif : Prise en compte du cahier des charges. Création d'items - Items sur support papier - Items sur support numérique Expérimentation effectuée auprès d'un échantillon représentatif d'élèves.	Objectif : Analyse de l'expérimentation. Construction de l'évaluation à partir d'une sélection d'items - Items validés après l'expérimentation ; - Items d'ancrage (repris des évaluations précédentes). Evaluation effectuée auprès d'un échantillon représentatif d'élèves.	Objectif : Analyse des premiers résultats Publications : - Note d'information - Repères et références statistiques (RERS) - Etat de l'école

1.3 Construction du test

Le bureau de l'évaluation des élèves de la DEPP élabore des évaluations par disciplines et niveaux scolaires. La préparation des unités et de leurs constituants fait intervenir des concepteurs, généralement des enseignants. La coordination est assurée par un chef de projet, membre de l'équipe du bureau de l'évaluation des élèves. Une application dédiée leur permet de créer, modifier ou éditer leur unité ; en outre cette application permet au chargé d'étude de gérer l'ensemble de l'évaluation (cf. plus loin l'encadré « GEODE »).

1.3.1 Elaboration des items

Les items sont le fruit d'un travail collectif des concepteurs encadrés par le chargé d'études. Un item proposé par un concepteur, pédagogue de terrain ayant une bonne connaissance des pratiques de classe, fait l'objet d'une discussion jusqu'à aboutir à un consensus. L'item est alors validé par le chargé d'études. Ce questionnement fait ensuite l'objet d'un cobayage, c'est-à-dire d'une passation auprès d'une ou plusieurs classe(s) pour estimer sa difficulté, les durées de passation minimale, maximale et moyenne, et recueillir les réactions éventuelles des élèves. Un équilibre de proportion est recherché entre les items considérés comme étant "faciles", "moyennement faciles" ou "difficiles". Afin d'assurer une comparabilité, 108 items sur les 195 proposés aux élèves sont des items d'ancrage, autrement dit repris à l'identique des passations antérieures de 2006 et/ou de 2012. Deux

types de formats de questions sont utilisés : les questions fermées (QCM, QCM-images, série, série-images) et les questions ouvertes appelant une réponse écrite (réponse courte - un mot ou un groupe de mots - ou réponse longue - de type rédaction). Un entraînement est prévu au début de chaque cahier afin de familiariser les élèves avec le type de question rencontré. Les réponses des formats QCM, QCM-images, série (de type vrai/faux) et série-images sont saisies de manière automatisée à la fin de la passation. Les items aux formats série et série-images attendent chacun plusieurs réponses de l'élève, pour chacune des propositions ou sous-items. Ces réponses aux différentes propositions relevant d'un même item font alors l'objet d'un regroupement, après définition d'un seuil de validation avec le groupe de concepteurs. Autrement dit, selon le niveau de difficulté voulu pour l'item concerné, on détermine pour sa validation, au cas par cas, qu'il faut des réponses correctes pour l'ensemble des propositions ou sous-items, ou pour un nombre précis d'entre eux. Pour ce qui est des items d'ancrage, repris des passations antérieures, ce seuil reste identique. Les réponses des formats " réponse libre de l'élève " sont corrigées par des experts. Cela suppose la mise en place d'un dispositif de corrections, nécessitant la formation technique des correcteurs et l'élaboration d'un cahier de correction précis, déclinant les attendus pour éviter toute subjectivité ou la validation de réponses trop imprécises ou trop succinctes. Ce dispositif de correction à distance s'appuie sur le logiciel AGATE (cf. partie " Analyse des items ").

GEODE (Gestion électronique d'outils et documents d'évaluation) : un outil de création et de stockage des évaluations

Objectifs

Le bureau de l'évaluation des élèves coordonne chaque année plusieurs évaluations afin d'apprécier le niveau de connaissances et de compétences des élèves en référence aux programmes officiels. Ces évaluations utilisent des livrets d'évaluation sur format papier et/ou électroniques.

L'application GEODE (gestion électronique d'outils et documents d'évaluation) est une application de création et de gestion dématérialisées des évaluations. Développée en 2009, elle a pour objectif de soutenir de bout en bout le processus de création des exercices et de constitution des cahiers et supports électroniques, allant jusqu'au bon à imprimer pour les évaluations papiers ou la génération d'une maquette de site web pour l'évaluation électronique.

L'application permet la conservation, l'indexation et la recherche des docu-

ments ou fichiers joints. Une partie des données textuelles, images, sons ou vidéos y est donc stockée que ce soit pour les évaluations papiers (cahier d'évaluations) ou les évaluations électroniques (outil de maquettage).

Principes fonctionnels

GEODE permet ainsi l'harmonisation des pratiques et formats de documents. La dématérialisation des documents rend indépendant l'éditeur (OpenOffice, Word,...) tout en permettant des variantes selon les disciplines. L'application dispose d'une GED (gestion électronique de documents) intégrée capable de gérer du texte, des images, du son et de la vidéo sous forme d'objets. Les cahiers sont générés au format Open Office principalement pour le format « papier », l'utilisation de la même technologie permet de générer du HTML pour la partie évaluation électronique (outil de maquettage).

1.3.2 Constitution des cahiers

Afin de pouvoir évaluer un nombre important d'items sans allonger le temps de passation pour l'élève, CEDRE utilise la méthodologie des cahiers tournants. Les items sont ainsi répartis dans des blocs d'une durée de 25 minutes et les blocs sont ensuite distribués dans les cahiers tout en respectant certaines contraintes : chaque bloc doit se retrouver un même nombre de fois au total et chaque association de blocs doit figurer au moins une fois dans un cahier. Ce dispositif, couramment utilisé dans les évaluations bilans, notamment les évaluations internationales, permet d'estimer la probabilité de réussite de chaque élève à chaque item sans qu'il ait à répondre à l'ensemble de ceux-ci. Au final, pour l'évaluation CEDRE histoire-géographie 2017, chaque cahier comprend deux séquences de 50 minutes, obligatoirement espacées d'une pause (au moins le temps d'une récréation). La seconde séance se termine par un questionnaire de contexte, d'une durée de 10 minutes environ, identique dans tous les cahiers, dans lequel l'élève doit répondre à des questions concernant notamment l'environnement scolaire, son intérêt et sa motivation pour l'histoire, la géographie et l'enseignement moral et civique. L'anonymat des élèves et des personnels est respecté, chaque cahier étant repéré par un numéro. Une fois l'évaluation terminée, les cahiers et questionnaires sont renvoyés dans des conditionnements prévus à cet effet, préaffranchis et pré-étiquetés. Aucun travail de correction n'a été demandé aux établissements.

L'évaluation CEDRE 2017 est constituée de 13 cahiers tournants intégrant un ensemble de 13 blocs d'évaluations contenant des items de 2006 et 2012 repris à l'identique pour assurer une comparaison diachronique et de nouveaux items qui ont fait l'objet d'une expérimentation en 2016. Pour garantir la qualité de la

comparaison avec 2006 et 2012, notamment en termes de passation des épreuves, l'évaluation de 2017 s'appuie sur 195 items dont 108 d'ancrage soit 55 %.

Tableau 5 – Exemple de répartition des blocs dans les cahiers

Cahier	Bloc 1	Bloc 2	Bloc 3	Bloc 4
E01	B5	B6	B12	B7
E02	B4	B13	B3	B8
E03	B6	B3	B2	B9
E04	B12	B2	B1	B13
E05	B3	B1	B7	B11
E06	B2	B7	B8	B10
E07	B1	B8	B9	B5
E08	B7	B9	B13	B4
E09	B8	B13	B11	B6
E10	B9	B11	B10	B12
E11	B13	B10	B5	B3
E12	B11	B5	B4	B2
E13	B10	B4	B6	B1

1.4 Passation des évaluations

La passation de l'évaluation finale a eu lieu en mai 2017. Comme en 2012, cette évaluation a été précédée d'une expérimentation l'année N- 1 de façon à tester un grand nombre d'items auprès d'un échantillon réduit d'établissements. Dans chaque établissement, une personne a été désignée comme étant l'administrateur du test, son rôle étant de veiller au strict respect de la procédure à suivre pour que l'évaluation soit passée dans les meilleures conditions, quel que soit l'établissement ; la collecte de l'information s'est faite par questionnaires "papier-crayon".

L'anonymat des élèves et des personnels est respecté, chaque cahier étant repéré par un numéro. Une fois l'évaluation terminée, les cahiers et questionnaires sont renvoyés dans des conditionnements prévus à cet effet, préaffranchis et pré-étiquetés. Aucun travail de correction n'a été demandé aux établissements.

2 Sondage

2.1 Méthodes

2.1.1 Tirage équilibré de classes de 3e

De manière générale, pour le secondaire, deux options de tirage peuvent être considérées : soit un sondage par grappe en sélectionnant un échantillon de classes et tous les élèves des classes tirées au sort participent à l'évaluation ; soit un premier degré qui concerne les établissements puis un second degré où un nombre d'élèves fixe dans chaque établissement est sélectionné [1]. Les évaluations CEDRE suivent la première option tandis que l'évaluation PISA suit la seconde. Des simulations ont permis de montrer que les niveaux de précision des deux options sont très proches, dès lors que le tirage est équilibré (cf. encadré « Tirage d'établissement *versus* tirage de classes »). Le choix de sondages par grappe est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'une classe au collège permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les établissements tirés.

On note U la population visée par une évaluation donnée, Y la variable d'intérêt (typiquement le score à l'évaluation, ou bien une indicatrice de difficulté), X une variable auxiliaire, c'est-à-dire connue pour l'ensemble des élèves de la population U. Un échantillon S d'élèves est sélectionné dans la population U. Chaque élève i a la probabilité π_i d'être sélectionné dans l'échantillon S (probabilité d'inclusion). Enfin, les poids de sondages, définis comme les inverses des probabilités d'inclusion π_i, sont notés d_i.

Un échantillon équilibré est un échantillon qui est représentatif de la population au regard de certaines variables auxiliaires. Cela signifie que dans un échantillon équilibré, l'estimateur du total d'une variable auxiliaire X sera exactement égal au vrai total de la variable X dans la population.

Cette propriété s'écrit :

$$\sum_{i \in S} \frac{X_i}{\pi_i} = \sum_{i \in U} X_i \tag{1}$$

1. Dans ce second cas, les établissements sont tirés proportionnellement à leur taille (nombre d'élèves). En effet, une fois que les établissements sont échantillonnés, un nombre fixe d'élèves est alors sélectionné quel que soit l'établissement. Par conséquent, les élèves des grands établissements ont moins de chance d'être tirés au sort que les élèves des petits établissements. Le tirage proportionnel à la taille permet ainsi de rétablir l'égalité des probabilités de tirage.

Tirage d'établissements *versus* Tirage de classes

Pour faciliter la logistique dans les collèges, nous réalisons un tirage de classes de 3e, puis tous les élèves de la classe sélectionnée passent l'évaluation. On peut donc s'interroger sur la perte de la précision liée à cet effet de grappe.
Pour comparer la précision entre un tirage d'établissement et un tirage de classes, nous avons réalisé des simulations à partir de la base des notes au brevet en 2009 (Garcia, Le Cam, & Rocher, 2015).
Nous avons comparé deux stratégies d'échantillonnage. Il s'agit à chaque fois d'échantillons stratifiés à deux degrés :

- Tirage équilibré d'établissement puis tirage de 30 élèves dans chaque établissement sélectionné ;
- Tirage équilibré de classe puis sélection de tous les élèves des classes sélectionnées.

La stratification a été effectuée selon le secteur d'enseignement et dans chaque strate 2 000 élèves ont été échantillonnés.
Pour chacune des deux stratégies, 1 000 échantillons ont été tirés. Puis on calcule la moyenne des erreurs standards des notes moyennes en français, mathématiques et histoire-géographie. Le tableau ci-dessous montre que les deux stratégies de tirage ont des niveaux équivalents de précision.

Comparaison des erreurs standards (Garcia et al., 2015)

	Echantillon équilibré d'établissements	Echantillon équilibré de classes
Français	0,07	0,07
Mathématiques	0,11	0,11
Histoire-Géographie	0,08	0,08

Les échantillons équilibrés ont donc comme propriété de fournir une photographie parfaite de la population, au regard des variables auxiliaires connues, ce que ne garantit pas une procédure aléatoire simple d'échantillonnage. En théorie, ils permettent également d'améliorer la précision des estimateurs s'il existe un lien entre la variable d'intérêt et les variables auxiliaires.

Le tirage équilibré est réalisé grâce au programme CUBE développé par l'INSEE et mis à disposition sous forme de macro SAS. La documentation complète est disponible sur le site Internet de l'INSEE (Rousseau & Tardieu, 2004). L'algorithme permet de choisir de manière aléatoire un échantillon parmi tous

les échantillons possibles respectant les contraintes reposant sur les variables auxiliaires. Il se déroule en deux phases : une « phase de vol » et une « phase d'atterrissage ». Durant la phase de vol, toutes les contraintes sont respectées. Elle se termine si un échantillon équilibré de manière parfaite est trouvé ou s'il n'est pas possible de trouver un échantillon en respectant toutes les contraintes. Si la phase de vol n'a pas abouti à un échantillon, la phase d'atterrissage débute. Elle consiste au relâchement des contraintes et au choix optimal de l'échantillon selon le critère choisi par l'utilisateur (ordre de priorité sur les contraintes, relâchement de la contrainte avec un coût minimal sur l'équilibrage ou garantie d'un échantillon de taille fixe).

Par ailleurs, au moment du tirage de l'échantillon, les collèges dont une classe a déjà été sélectionnée pour une autre évaluation la même année sont exclus de la base de sondage. Les probabilités d'inclusion sont donc recalculées pour tenir compte de ces exclusions tout en gardant une représentativité nationale (cf. encadré « tirage équilibré après élimination de la base des échantillons précédemment tirés »).

2.1.2 Redressement de la non réponse : calage sur marges

Comme toute enquête réalisée par sondage, les évaluations des élèves sont exposées à la non-réponse. Bien que les taux de retour soient élevés, il est nécessaire de tenir compte de la non-réponse dans les estimations car celle-ci n'est pas purement aléatoire (par exemple, la non-réponse est plus élevée chez les élèves en retard). Afin de la prendre en compte, un calage sur marges est effectué à l'aide de la macro CALMAR, également disponible sur le site Internet de l'INSEE. La méthode de calage sur marges consiste à modifier les poids de sondage d_i des répondants de manière à ce que l'échantillon ainsi repondéré soit représentatif de certaines variables auxiliaires dont on connait les totaux sur la population (Sautory, 1993). C'est une méthode qui permet de corriger la non-réponse mais également d'améliorer la précision des estimateurs. En outre, elle a pour avantage de rendre cohérents les résultats observés sur l'échantillon pour ce qui concerne des informations connues sur l'ensemble de la population.

Les nouveaux poids w_i, calculés sur l'échantillon des répondants S', vérifient l'équation suivante pour les K variables auxiliaires sur lesquelles porte le calage :

$$\forall k = 1...K, \sum_{i \in S'} w_i X_i^k = \sum_{i \in U} X_i^k \tag{2}$$

Ils sont obtenus par minimisation de l'expression $\sum_{i \in S'} d_i G(\frac{w_i}{d_i})$ où G désigne une fonction de distance, sous les contraintes définies dans l'équation 2.

Tirage équilibré après élimination de la base des échantillons précédemment tirés

La situation est la suivante : un échantillon d'établissements a été sélectionné pour participer à une évaluation ; un deuxième échantillon doit être tiré pour une autre évaluation. Nous souhaitons éviter que des établissements soient interrogés deux fois. Il s'agit donc de gérer le non-recouvrement entre les échantillons et d'assurer également un tirage équilibré du deuxième échantillon. Nous nous concentrons ici sur le non-recouvrement des échantillons mais notons qu'une approche plus générale incluant un taux de recouvrement non nul (pour permettre des analyses croisées entre enquêtes) est en cours de développement avec une application à des données issues d'évaluations standardisées (Christine & Rocher, 2012).

Formulation du problème et notations

Un échantillon S_1 a été tiré. Il est connu et les probabilités d'inclusion des établissements π_j^1 sont également connues. On souhaite alors tirer un échantillon S_2 dans la population U avec les probabilités π_j^2, mais sans aucun recouvrement avec l'échantillon S_1. On va donc tirer l'échantillon S_2 dans la population $U(S_1)$, c'est-à-dire la population U privée des établissements de l'échantillon S_1 qui appartiennent à U. Notons d'emblée que S_1 n'a pas nécessairement été tiré dans U, mais potentiellement dans une autre population, plus large ou plus réduite ; cela n'affecte en rien la formulation envisagée ici. Notons également que l'indice j est utilisé ici : il concerne les établissements et non les élèves, représentés par l'indice i.

Il s'agit donc de procéder à un tirage conditionnel. On note π_j^{2/S_1} les probabilités d'inclusion conditionnelles des établissements dans le second échantillon S_2, sachant que le premier échantillon est connu. Ces probabilités conditionnelles peuvent s'écrire :

$$\pi_j^{2/S_1} = \begin{cases} \lambda_j & \text{si } j \notin S_1 \\ 0 & \text{si } j \in S_1 \end{cases} \text{, avec } \lambda_j \in [0,1]$$

On a $\pi_j^2 = E(\pi_j^{2/S_1}) = \lambda_j(1-\pi_j^1)$ d'où $\lambda_j = \frac{\pi_j^2}{1-\pi_j^1}$

Équilibrage

On souhaite maintenant que l'échantillon S_2 soit équilibré selon certaines

variables (nombre d'élèves en retard, etc.). Soit X une variable d'équilibrage, la condition s'écrit :

$$\sum_{j \in S_2} \frac{X_j}{\pi_j^2} = \sum_{j \in U} X_j$$

Pour arriver à ce résultat, le principe est de tirer S_2 dans $U(S_1)$ avec les probabilités d'inclusion λ_j et avec une condition d'équilibrage sur la variable $X_j/(1-\pi_j^1)$.

Ainsi, on aura :

$$\sum_{j \in S_2} \frac{X_j}{\pi_j^2} = \sum_{j \in S_2} \frac{X_j}{\lambda_j(1-\pi_j^1)} = \sum_{j \in U(S_1)} \frac{X_j}{1-\pi_j^1}$$

Or, en espérance on a

$$E(\sum_{j \in U(S_1)} \frac{X_j}{1-\pi_j^1}) = E(\sum_{j \in U} \frac{X_j}{1-\pi_j^1} I_{j \notin S_1}) = \sum_{j \in U} X_j$$

La condition d'équilibrage initiale est donc remplie.

Condition fondamentale

Comme il s'agit d'une probabilité, la condition fondamentale est que $\lambda_j \in [0,1]$. Comme $\lambda_j = \frac{\pi_j^2}{1-\pi_j^1}$, la condition est en fait que

$$\pi_j^1 + \pi_j^2 \leq 1$$

Dans certains cas, par exemple des strates souvent sur-représentées comme les établissements situés dans des zones spécifiques concernant peu d'élèves (ex : REP+), cette condition pourrait ne pas être satisfaite. Cependant, de façon concrète, la condition a toujours été respectée dans les plans de sondage réalisés.

2.1.3 Calcul de précision : méthode

Les résultats des évaluations sont soumis à une variabilité qui dépend notamment des erreurs d'échantillonnage. Il est possible d'estimer statistiquement ces erreurs d'échantillonnage, appelées erreurs standard.

On note Y la variable d'intérêt (typiquement le score obtenu à une évaluation) et $\hat{\bar{Y}}$ l'estimateur de la moyenne de Y, qui constitue un estimateur essentiel sur lequel nous insistons dans la suite, bien que d'autres soient également au centre des analyses, comme ceux concernant la dispersion. La méthode retenue est cependant applicable à différents types d'estimateurs.

Nous souhaitons estimer la variance de cet estimateur, c'est-à-dire $V(\hat{\bar{Y}})$. En absence de formule théorique pour calculer $V(\hat{\bar{Y}})$, il existe plusieurs procédures permettant de l'estimer, c'est-à-dire de calculer $\hat{V}(\hat{\bar{Y}})$, l'estimateur de la variance d'échantillonnage. Il peut s'agir de méthodes de linéarisation des formules (Taylor) ou bien de méthodes empiriques (méthodes de réplication, jackknife, etc.). Ces méthodes sont bien décrites dans la littérature. Le lecteur est invité à consulter Tillé (2001) ou Ardilly (2006).

Cependant, lorsqu'un calage sur marges a été effectué, il faut en tenir compte pour le calcul de la précision. Dans ce cas, la variance de $\hat{\bar{Y}}$ est asymptotiquement équivalente à la variance des résidus de la régression de la variable d'intérêt sur les variables de calage.

En pratique, pour estimer la variance d'échantillonnage de $\hat{\bar{Y}}$, tenant compte du calage effectué, il convient alors d'appliquer la procédure suivante :

1. On effectue la régression linéaire de la variable d'intérêt sur les variables de calage, en pondérant par les poids initiaux. Les résidus e_i de cette régression sont calculés.
2. Les valeurs $g_i e_i$ sont calculées, où g_i représente le rapport entre les poids CALMAR (w_i) et les poids initiaux (d_i) : $g_i = \frac{w_i}{d_i}$
3. La variance d'échantillonnage de $\hat{\bar{Y}}$ est alors obtenue en calculant la variance d'échantillonnage de $g_i e_i$.

2.2 Echantillonnage

Champ

Le champ des évaluation CEDRE au collège est celui des élèves de 3e générale scolarisés dans des collèges publics et privés sous contrat de France métropolitaine.

La base de sondage utilisée est la base dite Scolarité construite par la DEPP. C'est une base de données individuelles anonymes contenant de nombreuses informations sur les élèves scolarisés une année scolaire donnée (date de naissance, PCS des parents, etc.). Nous disposons également d'informations sur les établissements scolaires, comme par exemple le secteur d'enseignement. Ces informations, qualifiées de variables auxiliaires, peuvent être utilisées au moment du tirage des échantillons, pour définir les variables de stratification. Préalablement au tirage, les établissements des échantillons d'autres opérations d'évaluations de la DEPP sont retirés de la base de sondage.

Stratification

Une stratification est réalisée en fonction du secteur d'enseignement :

1. Public hors éducation Prioritaire (PU)
2. Public en éducation prioritaire (EP)
3. Privé (PR)

Modalités de sélection

Le tirage est à deux degrés. Le premier degré de sondage est composé de classes (et non de collèges) tirées dans chaque strate avec allocation proportionelle. Le deuxième degré de sondage consiste à interroger tous les élèves de la classe sélectionnée (tirage par grappe).La macro CUBE de l'INSEE est utilisée pour garantir des échantillons équilibrés sur la base de sondage selon certaines variables

Dans chacune des 3 strates, le tirage est équilibré sur les variables suivantes :

— Le nombre total d'élèves de 3e
— L'indice de position sociale (Rocher, 2016)
— Le nombre d'élèves de 3e en retard dans la population
— Le nombre de garçons de 3e dans la population

Echantillon 2017

L'échantillon vise 6 000 élèves répartis proportionnellement selon les trois strates.

Base de sondage

Le tableau 6 présente les exclusions dans la population ciblée.

Tableau 6 – Exclusions pour la base de sondage - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

	Établissements	**Elèves**
Etab. accueillant des élèves	8 439	835 310
On retire les COM	8 400	830 918
On retire les étab hors contr	8 220	828 467
On retire les EREA	8 150	827 124
On retire les UPE2A	8 140	826 237
On retire les ULIS	8 124	823 869
On retire les DOM	7 865	784 876
On ne garde que les collèges	6 694	756 264
On ne garde que les 3ème générales	6 691	733 456
Base CEDRE 3e	6 691	733 456
On retire les échantillons PISA et ICILS	6 657	729 617
Base de tirage	**6 657**	**729 617**

Le tableau 7 présente la répartition de la population ciblée selon le secteur d'enseignement.

Tableau 7 – Répartition dans la base de sondage - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Strate	**Établissements**	**Élèves**
1. Public hors EP	4 106	472 262
2. EP	961	100 894
3. Privé	1 624	160 300
Total	**6 691**	**733 456**

Échantillon

Le tableau 8 présente la répartition de l'échantillon selon le secteur d'enseignement. Au total, 195 écoles ont été sélectionnées.

Tableau 8 – Répartition dans l'échantillon - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Strate	**Établissements**	**Élèves**
1. Public hors EP	122	3 247
2. EP	31	699
3. Privé	42	1 117
Total	**195**	**5 063**

2.3 État des lieux de la non-réponse

2.3.1 Non-réponse totale

Parmi la non-réponse totale, nous distinguons la non-réponse des établissements de la non-réponse des élèves des établissements participants. Les chiffres suivants ont été observés pour 2017.

96.9 % des établissements de l'échantillon ont répondu à l'évaluation (tableau 9). 89.7 % des éffectifs attendus ont participé (tableau 10).

Tableau 9 – Non-réponse des établissements - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Strate	Nb établissements attendus	Nb établissements répondants	% d'établissements répondants
1. Public hors EP	122	122	100 %
2. EP	31	27	87.1 %
3. Privé	42	40	95.2 %
Total	**195**	**189**	**96.9 %**

Tableau 10 – Non-réponse des élèves - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Strate	Nb élèves attendus	Nb élèves répondants	% d'élèves répondants
1. Public hors EP	3 247	2 967	91.4 %
2. EP	699	565	80.8 %
3. Privé	1 117	1 009	90.3 %
Total	**5 063**	**4 541**	**89.7 %**

2.3.2 Valeurs manquantes et imputation

Dans le cas où certaines données sont manquantes, nous procédons à des imputations. Cela concerne uniquement les variables sexe et année de naissance, afin de pouvoir réaliser des statistiques selon ces variables sur l'échantillon complet, quelle que soit l'analyse. Nous imputons aléatoirement les valeurs manquantes de ces deux variables, de manière à respecter la répartition des répondants.

2.3.3 Non-réponse partielle et terminale

Lorsque des non-réponses sont observées aux items, nous distinguons les cas suivants :

- La non-réponse partielle : un élève n'a pas répondu à certains items dans le cahier.
- La non-réponse terminale : un élève s'est arrêté avant la fin du cahier soit par manque de temps soit par abandon.

Dans le premier cas, les non-réponses sont traitées comme des échecs (code "0"). Le second cas conduit à déterminer des règles. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes sont donc traitées de manière structurelle (code "s"). La non réponse terminale a été étudiée par séquence et par cahier. Si un élève a passé moins de 50 % d'une séquence, on considère qu'il n'a pas vu la séquence (code "s").

Parmi les élèves concernés, la non-réponse terminale représente en moyenne :

- 21.3 items pour la séquence 1
- 23.7 items pour la séquence 2

On considère que :

- 120 élèves n'ont pas vu la séquence 1, dont :
  - 85 n'ont répondu à aucun items de la séquence
  - 35 ont répondu à moins de 50 % de la séquence
- 186 élèves n'ont pas vu la séquence 2, dont :
  - 133 n'ont répondu à aucun items de la séquence
  - 53 ont répondu à moins de 50 % de la séquence

Les élèves dont toutes les séquences sont codées en "s" sont classés en non réponse totale. C'est le cas pour 39 élèves.

2.4 Redressement

Pour tenir compte de la non réponse, l'échantillon a été redressé à l'aide d'un calage sur marge. Préalablement au calage, on effectue tout d'abord une post-

stratification. Puis, deux variables de calage sont utilisées :

— la répartition selon le sexe dans la population ;
— la répartition selon le retard scolaire.

Tableau 11 – Comparaison entre les marges de l'échantillon et les marges dans la population - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Modalité	Variable	Marge échantillon	Marge population	Pourcentage échantillon	Pourcentage population
Retard	1	96 979.86	109 144	13.22	14.88
	2	636 476.09	624 312	86.78	85.12
Sexe	1	364 890.26	366 835	49.75	50.01
	2	368 565.69	366 621	50.25	49.99
Strate	1	472 261.97	472 262	64.39	64.39
	2	100 894	100 894	13.76	13.76

2.5 Précision

L'erreur standard (se) peut être calculée sur le score moyen de chaque année (tableau 12).

Tableau 12 – Scores moyens et erreurs standard associées - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Année	Score moyen	Erreur standard
2006	250	1.95
2012	239.9	1.61
2017	245.2	1.24

Pour savoir par exemple si l'évolution entre 2012 et 2017 est significative , il faut calculer la valeur suivante :

$$\frac{\left|\hat{\bar{Y}}_{2017} - \hat{\bar{Y}}_{2012}\right|}{\sqrt{se^2_{\hat{\bar{Y}}_{2017}} + se^2_{\hat{\bar{Y}}_{2012}}}} \tag{3}$$

Entre 2012 et 2017, on obtient une valeur de 2.63 (supérieure à 1.96). Cela signifie que l'évolution du score moyen est statistiquement significative.

Les erreurs standards sont également calculées pour les répartitions dans les différents groupes de niveaux (tableaux 13 et 14).

Tableau 13 – Répartitions en % dans les groupes de niveaux - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2006	2.4	12.6	28.2	29.7	17.2	10
2012	3.6	17.3	30.3	27.6	14.7	6.7
2017	2.6	14.7	28.6	29.9	17	7.2

Tableau 14 – Erreurs standards des répartitions en % dans les groupes de niveaux - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Année	Groupe <1	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
2006	0.4	0.9	1.1	1.1	0.8	0.8
2012	0.3	0.9	0.9	0.8	0.7	0.6
2017	0.3	0.6	0.8	0.8	0.7	0.6

Design effect

L'effet du plan de sondage (*Design Effect*) permet de rapporter l'erreur de mesure faite par un tirage spécifique à l'erreur de mesure qui aurait été faite en procédant à un sondage aléatoire simple (SAS) du même nombre d'élèves. Pour la moyenne d'une variable Y et un plan de sondage complexe P :

$$D_{eff} = \frac{V_P(\hat{\bar{Y}})}{V_{SAS}(\hat{\bar{Y}})} \tag{4}$$

Tableau 15 – Effet du plan de sondage - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Année	Erreur Standard	Erreur SAS	*Design Effect*
2006	1.95	0.65	2.98
2012	1.61	0.73	2.2
2017	1.24	0.72	1.73

Dans le cas d'un sondage en grappes, la précision est dégradée en comparaison d'un sondage aléatoire simple. Cela signifie qu'en 2017, un sondage aléatoire simple avec un effectif 1.73 fois moins important aurait conduit au même niveau de précision.

3 Analyse des items

3.1 Méthodologie

Pour une description générale de la méthodologie psychométrique employée dans les évaluations standardisées de compétences des élèves, le lecteur est invité à consulter Rocher (2015).

3.1.1 Approche classique

Dans un premier temps, nous posons quelques notations et nous présentons les principales statistiques descriptives utilisées pour décrire un test, issues de la « théorie classique des tests » que nous évoquons rapidement.

Réussite et score

On note n le nombre d'élèves ayant passé une évaluation composée de J items. On note Y_i^j la réponse de l'élève i ($i = 1, ..., n$) à l'item j ($j = 1, ..., J$). Dans notre cas, les items sont dichotomiques, c'est-à-dire qu'ils ne prennent que deux modalités (la réussite ou l'échec) :

$$Y_i^j = \begin{cases} 1 & \text{si l'élève } i \text{ réussit l'item } j \\ 0 & \text{si l'élève } i \text{ échoue à l'item } j \end{cases} \tag{5}$$

Le taux de réussite à l'item j est la proportion d'élèves ayant réussi l'item j. Il est noté p_j :

$$p_j = \frac{1}{n} \sum_{i=1}^{n} Y_i^j \tag{6}$$

Le taux de réussite d'un item renvoie à son niveau de difficulté. C'est certainement la caractéristique la plus importante, qui permet de construire un test de niveau adapté à l'objectif de l'évaluation, en s'assurant que les différents niveaux de difficulté sont balayés.

Le score observé à l'évaluation pour l'élève i, noté S_i, correspond au nombre d'items réussis par l'individu i :

$$S_i = \sum_{j=1}^{J} Y_i^j \tag{7}$$

La théorie classique des tests a précisément pour objet d'étude le score S_i obtenu par un élève à un test. Elle postule notamment que ce score observé résulte de la somme d'un score « vrai » inobservé et d'une erreur de mesure. Un certain

nombre d'hypothèses portent alors sur le terme d'erreur (pour plus d'informations, cf. par exemple Laveault et Gregoire, 2002).

Fidélité

Dans le cadre de la théorie classique des tests, la fidélité (*reliability*) est définie comme la corrélation entre le score observé et le score vrai : le test est fidèle, lorsque l'erreur de mesure est réduite. Une manière d'estimer cette erreur de mesure consiste par exemple à calculer les corrélations entre les différents sous-scores possibles : plus ces corrélations sont élevées, plus le test est dit fidèle [2].

Le coefficient α de Cronbach est un indice destiné à mesurer la fidélité de l'épreuve. Il est compris entre 0 et 1. Sa version « standardisée » s'écrit :

$$\alpha = \frac{J\bar{r}}{1 + (J-1)\bar{r}} \tag{8}$$

où $\bar{r}$ est la moyenne des corrélations inter-items.

De ce point de vue, cet indicateur renseigne sur la consistance interne du test. En pratique, une valeur supérieure à 0,8 témoigne d'une bonne fidélité [3].

Indices de discrimination

Des indices importants concernent le pouvoir discriminant des items. Nous présentons ici l'indice « r-bis point » ou coefficient point-bisérial qui est le coefficient de corrélation linéaire entre la variable indicatrice de réussite à l'item Y^j et le score S.

Appelé également « corrélation item-test », il indique dans quelle mesure l'item s'inscrit dans la dimension générale. Une autre manière de l'envisager consiste à le formuler en fonction de la différence de performance constatée entre les élèves qui réussissent l'item et ceux qui l'échouent.

2. Notons au passage que la naissance des analyses factorielles est en lien avec ce sujet : Charles Spearman cherchait précisément à dégager un facteur général à partir de l'analyse des corrélations entre des scores obtenus à différents tests.

3. La littérature indique plutôt un seuil de 0,70 (Peterson, 1994). Cependant, comme le montre la formule ci-dessus, le coefficient α est lié au nombre d'items, qui est important dans les évaluations conduites par la DEPP afin de couvrir les nombreux éléments des programmes scolaires. Des facteurs de correction existent néanmoins et permettent de comparer des tests de longueur différentes.

En effet, on peut montrer que

$$r_{bis-point}(j) = corr(Y^j, S) = \frac{\bar{S}_{(j1)} - \bar{S}_{(j0)}}{\sigma_S}\sqrt{p_j(1-p_j)} \tag{9}$$

où $\bar{S}_{(j1)}$ est le score moyen sur l'ensemble de l'évaluation des élèves ayant réussi l'item j, $\bar{S}_{(j0)}$ celui des élèves l'ayant échoué et σ_S est l'écart-type des scores.

C'est donc bien un indice de discrimination, entre les élèves qui réussissent et ceux qui échouent à l'item. En pratique, on préfère s'appuyer sur les $r_{bis-point}$ corrigés, c'est à dire calculés par rapport au score à l'évaluation privée de l'item considéré. Une valeur inférieure à 0,2 indique un item peu discriminant (Laveault et Grégoire, 2002).

3.1.2 Analyse factorielle des items

L'analyse factorielle permet d'étudier la structure des données et, plus particulièrement, la structure des corrélations entre les variables observées (ou manifestes) [4]. Il s'agit d'identifier les différentes dimensions sous-jacentes aux réussites observées et surtout d'évaluer le poids de la dimension principale, dans la mesure où c'est une optique unidimensionnelle qui sera envisagée lors de la modélisation.

Dans le cas où les items sont dichotomiques, la matrice des corrélations entre items est en fait la matrice des coefficients ϕ, qui sont bornés selon les taux de réussite aux items (Rocher, 1999). Une analyse factorielle basée sur cette matrice peut donc montrer quelques faiblesses : des facteurs « artefactuels » sont susceptibles d'apparaître, en lien avec le niveau de difficulté des items et non avec les dimensions auxquelles ils se rapportent. De plus, d'un point de vue théorique, certaines hypothèses utiles pour l'estimation, comme la normalité des variables, ne sont pas envisageables.

L'optique retenue est alors de se ramener à un modèle linéaire : les variables observées catégorielles sont considérées comme la manifestation de variables latentes continues.

4. Notons qu'il s'agit ici d'analyse factorielle en facteurs communs et spécifiques et non d'analyse factorielle géométrique de type ACP ou ACM (pour des détails, consulter Rocher, 2013)

Les réponses à un item dichotomique sont définies de la manière suivante :

$$y_{ij} = \begin{cases} 0 & \text{si } z_{ij} \leq \tau_j \\ 1 & \text{si } z_{ij} > \tau_j \end{cases} \tag{10}$$

La réponse y_{ij} de l'élève i à l'item j est incorrecte tant que la variable latente Z_j reste en deça d'un certain seuil τ_j, qui dépend de l'item. Au-delà de ce seuil, la réponse est correcte.

L'analyse factorielle des items consiste donc en une analyse factorielle linéaire sur les variables continues Z_j. Deux modèles sont donc considérés. D'une part, une variable latente continue et conditionnant la réponse à l'item est fonction linéaire de facteurs communs et d'un facteur spécifique. D'autre part, un modèle de seuil représente la relation non linéaire entre la variable latente et la réponse à l'item. Ce procédé permet de se ramener à une analyse factorielle linéaire, à la différence que les variables Z_j ne sont pas connues. Il s'agit donc d'estimer la matrice de corrélation de ces variables, sous certaines hypothèses.

Considérons le lien entre deux items j et k. Si les variables latentes correspondantes Z^j et Z^k sont distribuées selon une loi normale bivariée, il est possible d'estimer le coefficient de corrélation linéaire de ces deux variables à partir du tableau croisant les deux items. C'est le coefficient de corrélation tétrachorique – ou polychorique dans le cas d'items polytomiques. L'estimation de ce coefficient par le maximum de vraisemblance requiert la résolution d'une double intégrale (pour les détails de l'estimation pour deux items dichotomiques, cf. Rocher, 1999). Pour plus de deux items, il devient difficile d'estimer de la même manière les coefficients de corrélation à partir de la distribution conjointe des items qui est une loi normale multivariée. C'est pourquoi les coefficients de corrélation tétrachorique sont estimés séparément pour chaque couple d'items. Ce procédé a le désavantage de conduire à une matrice de covariances qui n'est pas nécessairement semi-définie positive, donc potentiellement non inversible.

3.2 Codage des réponses aux items

3.2.1 Valeurs manquantes

Trois types de valeurs manquantes sont distinguées :

- Valeurs manquantes structurelles : l'élève n'a pas vu l'item. C'est le cas pour les cahiers tournants, où les élèves ne voient pas tous les items. Dans ce cas, on considère l'item comme *non administré*, l'absence de réponse n'est alors pas considérée comme une erreur.
- Absence de réponse : l'élève a vu l'item mais n'y a pas répondu. L'absence de réponse est alors considérée comme une erreur de la part de l'élève.

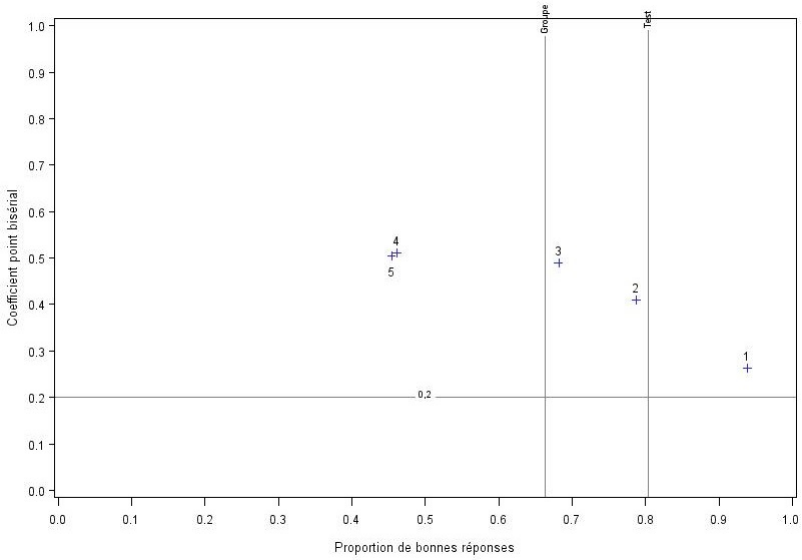
— Non-réponse terminale : l'élève s'est arrêté au cours de l'épreuve, potentiellement en raison d'un manque de temps. Des choix sont effectués pour déterminer le traitement de ces valeurs. Nous considérons que si un élève a passé moins de 50 % d'une séquence, il n'a pas vu la séquence, les valeurs manquantes sont alors traitées de manière structurelle. Sinon, elles sont traitées comme des échecs.

3.2.2 Regroupement des items

Les séries d'items comportant seulement deux réponses, comme les Vrai/Faux, font l'objet d'un traitement spécifique. Les items de ce type sont regroupés pour former un seul item à réponse binaire (réussite ou échec). En effet, la plus forte potentialité de réponse au hasard et l'inter-dépendance des items fragilisent leur utilisation individuelle.

Le regroupement de ces items consiste à faire la somme des indicatrices de réussite et à déterminer un seuil de maîtrise. Une visualisation graphique est utilisée pour fixer les scores « seuils » (cf. figure 1). Ce graphique représente le taux de réussite pour chaque seuil possible en fonction de la discrimination obtenu pour le seuil. Il permet de choisir la combinaison la mieux adaptée. Le score seuil doit préserver la discrimination de l'item regroupé et la difficulté peut être modulée en fonction des objectifs.

Figure 1 – Représentation graphique utilisée pour le regroupement d'items



Note de lecture : L'item présenté ici est une série de cinq questions de type « Vrai/Faux ». Chaque croix représente l'item correspondant au seuil de réussite retenu. Par exemple, si la réussite à l'ensemble est attribuée dès lors qu'une seule question est réussie, l'item obtenu a un taux de réussite d'environ 95 % et un coefficient bisérial d'environ 0,26. Si le seuil de réussite est fixé à 3 questions réussies sur 5, alors le taux de réussite baisse mécaniquement (autour de 65 % qui est le taux de réussite obtenu à l'ensemble des questions de cet item).

3.2.3 Traitement des données et correction des questions ouvertes

Tous les cahiers recueillis dans le cadre de cette opération ont été scannés par une société extérieure. Les réponses aux questions à choix multiples ainsi que les grilles d'évaluation remplies par les professeurs lors des séquences de travaux pratiques ont été numérisées et les codes de réponses stockés dans un fichier. En ce qui concerne les questions ouvertes, demandant une rédaction plus ou moins longue de la part des élèves (explication, schématisation...), elles ont été découpées en « imagettes » puis transmises au ministère afin d'être intégrées dans un logiciel de correction à distance (cf. encadré « AGATE »). Celui-ci nécessite la formation technique des correcteurs et l'élaboration d'un cahier des charges strict de corrections pour limiter la subjectivité des corrections. Une fois la correction terminée, les codes saisis par les correcteurs ont été stockés dans un fichier puis associés à ceux issus des réponses aux QCM.

AGATE : un outil de correction à distance des questions ouvertes

Objectifs

Le logiciel AGATE, qui a été développé par les informaticiens de la DEPP, permet une correction à distance des questions ouvertes. Le principe général du logiciel est de soumettre un lot d'imagettes (image scannée de la réponse d'un élève) à un groupe de correcteurs tout en paramétrant des contraintes de double correction et/ou d'auto-correction. Lorsque deux correcteurs corrigent la même imagette, il arrive parfois qu'il y ait une différence de codage. Cette imagette est alors proposée au superviseur qui arbitre et valide l'un des deux codages. Ce jeu de codages multiples incrémente des compteurs (temps de connexion, avancement général et taux d'erreur) qui sont autant d'indicateurs pour suivre la correction. A noter qu'un processus de déconnexion automatique d'un correcteur existe si le superviseur se rend compte d'un trop grand nombre d'erreurs de correction. Ce logiciel est utilisé depuis 2004 par le bureau des évaluations de la DEPP. Il a permis d'intégrer des questions ouvertes dans des évaluations à grandes échelles, aussi bien aux évaluations nationales qu'aux évaluations internationales telles PISA, TIMSS ou PIRLS. Les correcteurs n'ont plus à manipuler un nombre très important de cahiers et peuvent travailler de manière autonome lorsqu'ils le souhaitent, tout en maintenant un contact entre eux et les responsables de l'évaluation afin d'assurer une meilleure fiabilité de la correction.

Principes fonctionnels

Le chef de projet paramètre la session de correction. Il définit les groupes de correcteurs et supervise chaque groupe. Il intègre et vérifie les items mis en correction et ajuste les paramètres de double correction. Son rôle consiste également à répondre aux questions des correcteurs par le biais d'une messagerie intégrée au logiciel et à communiquer sa réponse également aux autres correcteurs. Le superviseur gère son groupe de correcteurs. Il anime la session de formation, qui consiste d'une part à communiquer aux télécorrecteurs une grille de correction très précises et d'autre part à corriger collectivement à blanc un nombre défini d'imagettes pour s'assurer de la compréhension et de la bonne mise en oeuvre des consignes. Puis, pendant la télécorrection, il arbitre les litiges lors des doubles-corrections. Le correcteur corrige les items en portant un codage de réussite/erreur sur chaque item. En cas de doute, il peut se référer à son superviseur de groupe. Une messagerie interne complète le dispositif et permet un échange de point de vue entre les différents acteurs.

3.3 Résultats

3.3.1 Pouvoir discriminant des items

Le calcul des indices de discrimination conduit à éliminer 35 items dont les indices *rbis-point* sont trop faibles :

- 12 items de 2006
- 6 items communs à 2006 et 2012
- 2 items communs à 2006, 2012 et 2017
- 7 items de 2012
- 3 items communs à 2012 et 2017
- 5 items de 2017

4 Modélisation

4.1 Méthodologie

4.1.1 Modèle de réponse à l'item

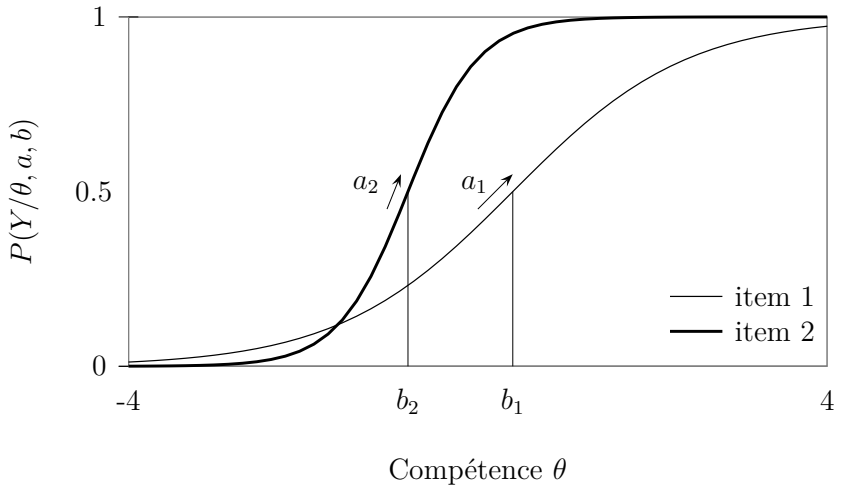
Le modèle de mesure utilisé est un modèle de réponse à l'item à deux paramètres avec une fonction de lien logistique (MRI 2PL) :

$$P_{ij} = P(Y_i^j = 1 \,|\, \theta_i, a_j, b_j) = \frac{e^{1,7a_j(\theta_i - b_j)}}{1 + e^{1,7a_j(\theta_i - b_j)}} \tag{11}$$

où la probabilité P_{ij} que l'élève i réussisse l'item j est fonction du niveau de compétence θ_i de l'élève i, du niveau de difficulté b_j de l'item j, ainsi que de la discrimination de l'item a_j ($a_j > 0$). La constante 1,7 est introduite pour rapprocher la fonction sigmoïde de la fonction de répartition de la loi normale.

La figure 2 représente les courbes caractéristiques de deux items selon cette modélisation.

Figure 2 – Modèle de réponse à l'item - 2 paramètres



Note de lecture : la probabilité de réussir l'item (en ordonnées) dépend du niveau de compétence (en abscisse). L'item 1 en trait fin est plus difficile que l'item 2 en trait plein ($b_1 > b_2$), et il est moins discriminant ($a_1 < a_2$).

L'avantage de ce type de modélisation, c'est de séparer deux concepts-clé, à savoir la difficulté de l'item et le niveau de compétence de l'élève. Les MRI ont un intérêt pratique pour la construction de tests et la comparaison entre différents groupes d'élèves : si le modèle est bien spécifié sur un échantillon donné, les paramètres des items – en particulier leurs difficultés – peuvent être considérés comme fixes et applicables à d'autres échantillons dont il sera alors possible de déduire les paramètres relatifs aux élèves – en particulier, leur niveau de compétence. Pour une présentation générale, le lecteur est invité à consulter Rocher (2015).

Autre avantage : le niveau de compétence des élèves et la difficulté des items sont placés sur la même échelle, par le simple fait de la soustraction $(\theta_i - b_j)$. Cette propriété permet d'interpréter le niveau de difficulté des items par rapprochement avec le continuum de compétence. Ainsi, les élèves situés à un niveau de compétence égal à b_j auront 50 % de chances de réussir l'item, ce que traduit visuellement la représentation des courbes caractéristiques des items (CCI) selon ce modèle (figure 2).

4.1.2 Procédures d'estimation

L'estimation est conduite en deux temps : l'estimation des paramètres des items puis l'estimation des θ en considérant les paramètres des items comme fixes. Nous donnons ici des éléments concernant ces procédures.

Estimation des paramètres des items

Nous reprenons les notations de l'équation (11) qui formule la probabilité P_{ij} d'un élève i de répondre correctement à un item j dans le cadre d'un modèle de réponse à l'item, avec les items sont dichotomiques.

Notons tout d'abord que les modèles présentés ne sont pas identifiables. En effet, les transformations $\theta_i^* = A\theta_i + B$, $b_j^* = Ab_j + B$ et $a_j^* = a_j/A$ avec A et B deux constantes $(A > 0)$, conduisent aux mêmes valeurs des probabilités. Dans CEDRE, nous levons l'indétermination en standardisant la distribution des θ pour les données du premier cycle (en l'occurrence, moyenne de 250 et écart-type de 50 pour l'année 2006).

Sous l'hypothèse d'indépendance locale des items [5], la fonction de vraisemblance s'écrit :

$$L(\mathbf{y}, \xi, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{J} P_{ij}^{y_{ij}} [1 - P_{ij}]^{1-y_{ij}} \tag{12}$$

5. Cette hypothèse signifie que les indicatrices de réussite des items sont indépendantes, conditionnellement au niveau de compétence θ. A niveau de compétence égal, deux items donnés ne sont pas corrélés : seule la compétence θ explique la corrélation entre deux items. Cette hypothèse est ainsi liée à l'hypothèse d'unidimensionnalité de θ (cf, Rocher, 2013).

où $\mathbf{y}$ est le vecteur des réponses aux items (*pattern*), ξ est le vecteur des paramètres des items.

La procédure MML (*Marginal Maximum Likelihood*) est utilisée. Elle consiste à estimer les paramètres des items en supposant que les paramètres des individus sont issus d'une distribution fixée *a prori* (le plus souvent normale). La maximisation de vraisemblance est *marginale* dans le sens où les paramètres concernant les individus n'apparaissent plus dans la formule de vraisemblance.

Si θ est considérée comme une variable aléatoire de distribution connue, la probabilité inconditionnelle d'observer un *pattern* $\mathbf{y}_i$ donné peut s'écrire :

$$P(\mathbf{y} = \mathbf{y}_i) = \int_{-\infty}^{+\infty} P(\mathbf{y} = \mathbf{y}_i \,|\, \theta_i) g(\theta_i) d\theta_i \tag{13}$$

avec g la densité de θ.

L'objectif est alors de maximiser la fonction de vraisemblance :

$$L = \prod_{i=1}^{n} P(\mathbf{y} = \mathbf{y}_i) \tag{14}$$

Cependant, l'annulation des dérivées de L par rapport aux a_j et aux b_j conduit à résoudre un système d'équations relativement complexe et à procéder à des calculs d'intégrales qui peuvent s'avérer très coûteux en termes de temps de calcul.

La résolution de ces équations est classiquement réalisée grâce à l'algorithme EM (*Expectation-Maximization*) impliquant des approximations d'intégrales par points de quadrature. L'algorithme EM est théoriquement adapté dans le cas de valeurs manquantes. Le principe général est de calculer l'espérance conditionnelle de la vraisemblance des données complètes (incluant les valeurs manquantes) avec les valeurs des paramètres estimées à l'étape précédente, puis de maximiser cette espérance conditionnelle pour trouver les nouvelles valeurs des paramètres. Le calcul de l'espérance conditionnelle nécessite cependant de connaître (ou de supposer) la loi jointe des données complètes. Une version modifiée de l'algorithme considère dans notre cas le paramètre θ lui-même comme une donnée manquante. Pour plus de détails, le lecteur est invité à consulter Rocher (2013).

En outre, ce cadre d'estimation permet aisément de traiter des valeurs manquantes structurelles, par exemple dans le cas de cahiers tournants ou bien dans le cas de reprise partielle d'une évaluation.

Estimation des niveaux de compétence

Une fois les paramètres des items estimés, ils sont considérés comme fixes et il est possible d'estimer les θ_i, par exemple *via* la maximisation de la vraisemblance donnée par l'équation (12).

Cependant, l'estimateur du maximum de vraisemblance, noté $\theta_i^{(ML)}$, est biaisé : les propriétés classiques de l'estimateur selon la méthode du maximum de vraisemblance ne sont pas vérifiées puisque le nombre de paramètres augmente avec le nombre d'observations. Ce biais vaut :

$$B(\theta_i^{(ML)}) = \frac{-J}{2I^2} \tag{15}$$

avec

$I = \sum_{j=1}^{J} \frac{P_{ij}^{'2}}{P_{ij}(1-P_{ij})} = \sum_{j=1}^{J} a_j^2 P_{ij}(1-P_{ij})$

et

$J = \sum_{j=1}^{J} \frac{P_{ij}^{'} P_{ij}^{''}}{P_{ij}(1-P_{ij})} = \sum_{j=1}^{J} a_j^3 P_{ij}(1-P_{ij})$

Pour obtenir un estimateur non biaisé, Warm (1989) a proposé de maximiser une vraisemblance pondérée $w(\theta)L(\mathbf{y}, \mathbf{a}, \mathbf{b}, \theta)$, en choisissant $w(\theta)$ de manière à ce que l'annulation de la dérivée du logarithme de la vraisemblance pondérée revienne à résoudre l'équation suivante :

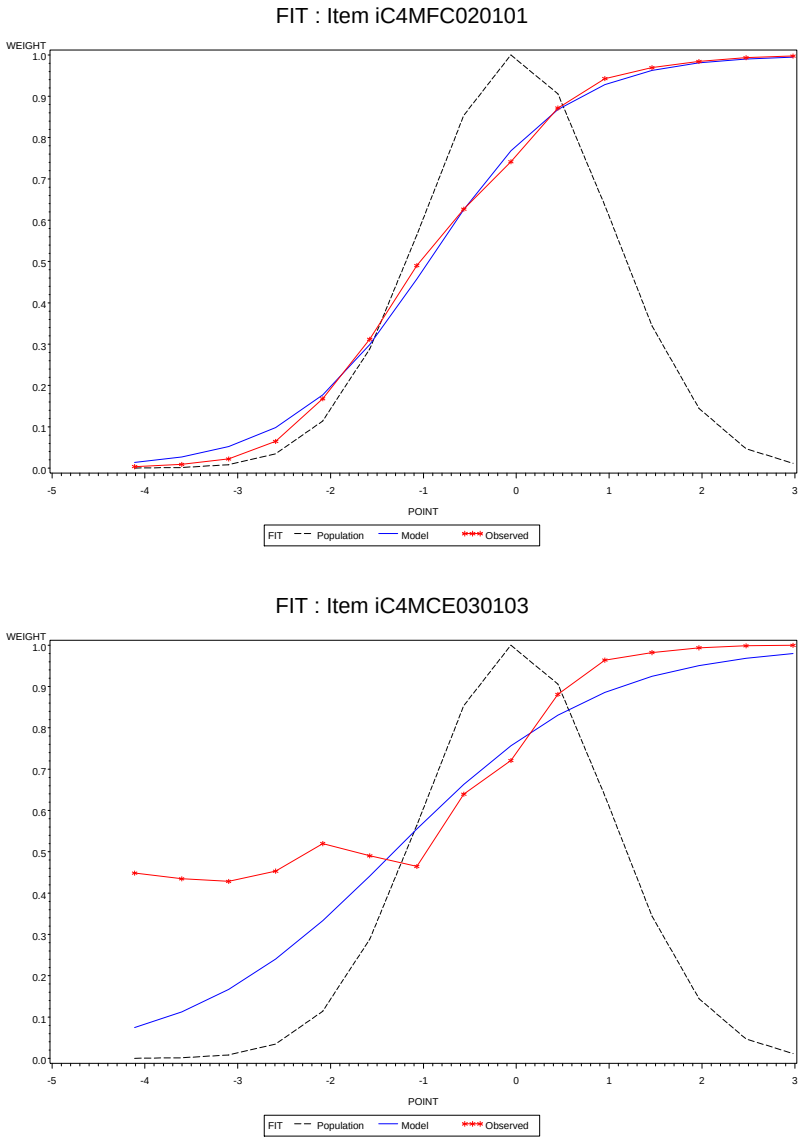
$$\frac{\partial lnL}{\partial \theta_i} + \frac{J}{2I} = 0 \tag{16}$$

4.1.3 Indice d'ajustement (FIT)

L'ajustement des items au modèle est étudié. Graphiquement, cela revient à comparer les courbes caractéristiques estimées avec les résultats observés (cf. figure 3). Certaines procédures proposent de comparer directement les probabilités théorique avec les proportions de réussite de groupes d'élèves. Plus généralement, nous pouvons écrire les résidus de la manière suivante :

$$z_{ij} = \frac{Y_i^j - P_{ij}}{\sqrt{P_{ij}(1-P_{ij})}} \tag{17}$$

Figure 3 – Exemples d'ajustements (FIT)



Note de lecture : La courbe bleue représente la courbe caractéristique de l'item telle qu'estimée par le modèle. La courbe en rouge relie des points qui correspondent aux taux de réussite observé à cet item pour 15 groupes d'élèves de niveaux de compétence croissants. Enfin, la courbe en pointillée représente la distribution des niveaux de compétence.
Clairement, l'ajustement du modèle est excellent pour l'item présenté en haut. Il est très mauvais pour celui du bas.

Les carrés des résidus suivent typiquement une loi du χ^2. L'indice *Infit* d'un item correspond à la moyenne pondérée des carrés des résidus, qui peut s'écrire :

$$Infit_j = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} w_{ij} z_{ij}^2 = \frac{1}{\sum_{i=1}^{n} w_{ij}} \sum_{i=1}^{n} (Y_i^j - P_{ij})^2 \tag{18}$$

avec le poids $w_{ij} = P_{ij}(1 - P_{ij})$. Une transformation de cet indice est utilisé de manière à obtenir une statistique suivant approximativement et empiriquement (le lien théorique n'est pas établi) une loi normale (Smith, Schumaker, & Bush, 1998).

4.1.4 Fonctionnement Différentiel d'Item (FDI)

Un fonctionnement différentiel d'item (FDI) apparaît entre des groupes d'individus dès lors qu'à niveau égal sur la variable latente mesurée, la probabilité de réussir un item donné n'est pas la même selon le groupe considéré. La question des FDI est importante car elle renvoie à la notion d'équité entre les groupes : un test ne doit pas risquer de favoriser un groupe par rapport à un autre.

Une définition formelle du FDI peut s'envisager à travers la propriété d'invariance conditionnelle : à niveau égal sur la compétence visée, la probabilité de réussir un item donné est la même quel que soit le groupe de sujets considéré. Formellement, un fonctionnement différentiel se traduit donc par :

$$P(Y \mid Z, G) \neq P(Y \mid Z) \tag{19}$$

où Y est le résultat d'une mesure de la compétence visée, typiquement la réponse à un item ; Z est un indicateur du niveau de compétence des sujets ; G est un indicateur de groupes de sujets.

Si la probabilité de réussite, conditionnellement au niveau mesuré, est différente selon les groupes d'élèves, alors il existe un fonctionnement différentiel.

En pratique, de très nombreuses méthodes ont été proposées afin d'identifier les FDI. Ces méthodes ont chacune des avantages en matière d'investigation des différents éléments pouvant conduire à l'apparition de ces FDI (Rocher, 2013). Dans le cas des évaluations standardisées menées à la DEPP, il s'agit avant tout d'identifier les fonctionnements différentiels pouvant apparaître entre deux moments de mesure, s'agissant des items repris à l'identique. Dans ce cas, les différentes méthodes d'identification donnent des résultats relativement proches.

Une stratégie très simple, employée dans CEDRE, consiste donc à comparer les paramètres de difficulté des items repris, estimés de façon séparée pour les deux

années. Si la difficulté d'un item a évolué, comparativement aux autres items, c'est le signe d'un fonctionnement différentiel, qui peut être lié par exemple à un changement de programmes ou de pratiques. Plus précisément, les paramètres des items sont estimés séparément pour les deux années, puis ajustés en tenant compte de la différence moyenne entre les deux séries de paramètres. La règle retenue pour identifier un FDI est celle d'un écart de paramètres de difficulté β d'au moins 0,5 (cf. Rocher, 2013 pour plus de détails).

4.1.5 L'information du test

Dans le cadre d'un modèle de réponse à l'item à deux paramètres, l'information d'un item j est définie par :

$$I_j(\theta) = (1,7a_j)^2 P_j(\theta)(1 - P_j(\theta)) \tag{20}$$

avec $P_j(\theta)$, la probabilité de réussite à l'item pour individu de compétence θ.

L'information moyenne du test pour un élève de compétence θ est la somme de l'information apporté par chaque item pour θ. La courbe d'information du test est tracée pour un ensemble de valeurs de θ. L'erreur de mesure étant inversement proportionnelle à l'information, cette courbe d'information permet de visualiser la précision avec laquelle le niveau de compétence θ des élèves est estimé.

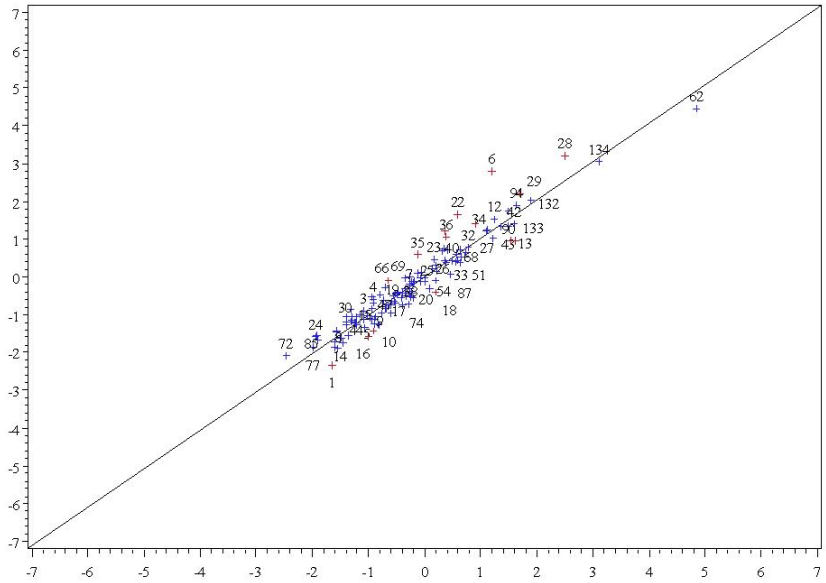
4.2 Résultats

4.2.1 Identification des fonctionnements différentiels d'items (FDI)

42 items ont été éliminés des calculs :

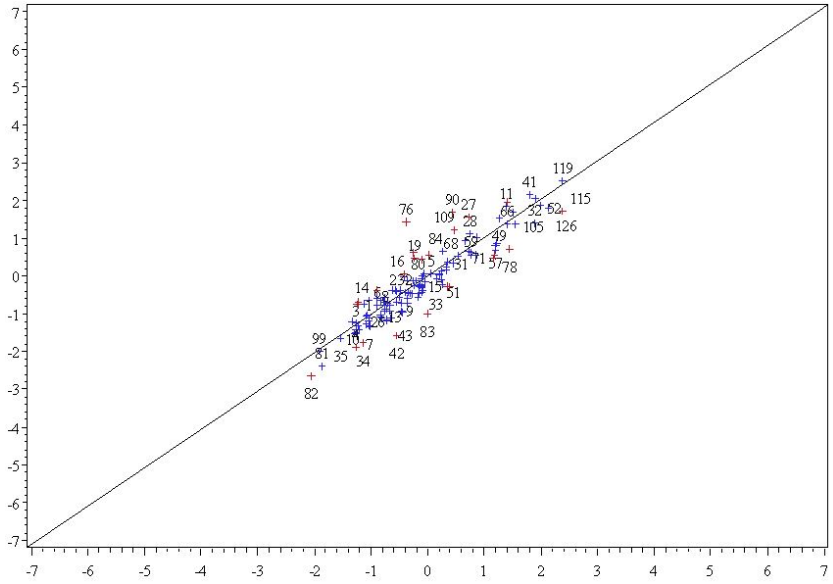
— 17 items pour 2006-2012
— 5 items pour 2012-2017
— 20 items pour 2006-2012-2017

Figure 4 – Comparaison des paramètres de difficulté 2006-2012 - (CEDRE HG 2017 Collège)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2006, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2012. Les items présentant un FDI apparaissent en bleu.

Figure 5 – Comparaison des paramètres de difficulté 2012-2017 - (CEDRE HG 2017 Collège)



Note de lecture : Les points sont les items. En abscisse figure la valeur des paramètres de difficulté estimés en 2012, et en ordonnée la la valeur des paramètres de difficulté estimés et ajustés pour l'année 2017. Les items présentant un FDI apparaissent en bleu.

4.2.2 Identification des items présentant un mauvais ajustement (FIT)

3 items ont été éliminés des calculs :

— 3 items pour 2006

4.2.3 Bilan de l'analyse des items

En considérant l'ensemble des items sur les 3 années, il y avait au départ :

— 235 items de 2006
— 13 items de 2012
— 92 items de 2017
— 138 items d'ancrage 2006-2012
— 43 items d'ancrage 2012-2017
— 95 items d'ancrage 2006-2012-2017

Cela représente 616 items passés par les élèves en tout, dont 230 en 2017.

Après suppression des items présentant un mauvais Rbis, un fonctionnement différentiel ou un mauvais ajustement, il reste :

— 220 items de 2006

— 6 items de 2012
— 87 items de 2017
— 115 items d'ancrage 2006-2012
— 35 items d'ancrage 2012-2017
— 73 items d'ancrage 2006-2012-2017

536 items sont donc conservés dans l'analyse, dont 195 utilisés dans l'évaluation 2017.

4.3 Calcul des scores

Comme indiqué précédemment, une analyse conjointe des données des 3 années a permis d'estimer les paramètres des items, puis les niveaux de compétences θ des élèves. Afin de lever l'indétermination du modèle, la moyenne des θ a été fixé à 250 et leur écart-type à 50, pour l'échantillon de 2006. Le tableau 16 présente les résultats obtenus.

Tableau 16 – Niveaux de compétences (moyennes des scores et écarts-types) - CEDRE 2017 Histoire-géographie et enseignement moral et civique Collège

Année	Score moyen	Écart-type
2006	250	50
2012	239.9	48.8
2017	245.2	48.2

5 Construction de l'échelle

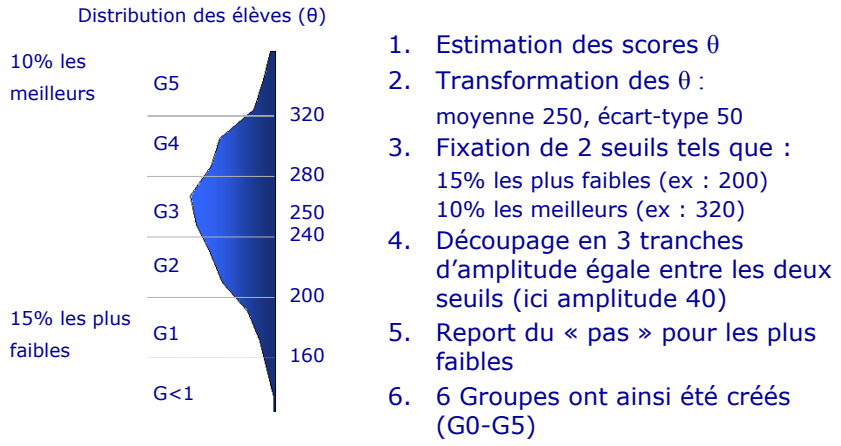
5.1 Méthode

Les modèles de réponse à l'item permettent de positionner sur une même échelle les paramètres de difficulté des items et les niveaux de compétences des élèves. Cette correspondance permet de caractériser les compétences maîtrisées pour différents groupes d'élèves.

Les scores en Histoire-géographie et enseignement moral et civique estimés selon le modèle de réponse à l'item présenté dans la partie précédente ont été standardisés de manière à obtenir une moyenne de 250 et un écart-type de 50 pour l'année 2006. Puis, comme le montre la figure 6, la distribution des scores est « découpée » en six groupes de la manière suivante : nous déterminons le score-seuil en-deça duquel se situent 15 % des élèves (groupes < 1 et 1), nous déterminons le score-seuil au-delà duquel se situent 10 % des élèves (groupe 5). Entre ces deux niveaux, l'échelle a été scindée en trois parties d'amplitudes de scores égales correspondant à trois groupes intermédiaires. Ces choix sont arbitraires et ont pour objectif de décrire plus précisément le continuum de compétence.

En effet, les modèles de réponse à l'item ont l'avantage de positionner sur la même échelle les scores des élèves et les difficultés des items. Ainsi, chaque item est associé à un des six groupes, en fonction des probabilités estimées de réussite selon les groupes. Un item est dit « maîtrisé » par un groupe dès lors que l'élève ayant le score le plus faible du groupe a au moins 50 % de chance de réussir l'item. Les élèves du groupe ont alors plus de 50 % de chance de réussir cet item.

Figure 6 – Principes de construction de l'échelle



5.2 Caractérisation des groupes de niveaux

A partir de cette correspondance entre les items et les groupes, une description qualitative et synthétique des compétences maîtrisées par les élèves des différents groupes est proposée.

Groupe < 1 (2,6 % des élèves)

Les élèves du groupe < 1 ne sont capables que de réponses ponctuelles et dispersées. Leurs problèmes de compréhension de l'écrit sur tous les supports accentuent encore leur difficulté à identifier des informations auxquelles ils ne peuvent le plus souvent donner sens.

Groupe 1 (14,7 % des élèves)

Les élèves du groupe 1 ont des connaissances et capacités très fragmentaires et restreintes. Leur compréhension du vocabulaire simple s'affirme cependant un peu par rapport au groupe < 1. Ils sont capables de prélever quelques informations très explicites sur des supports simples.

Groupe 2 (28,6 % des élèves)

Les élèves du groupe 2 ont un taux moyen de réussite à l'ensemble des items sensiblement plus élevé que celui du groupe 1. Ils restituent des connaissances correspondant surtout au programme de troisième. Ils en reconnaissent le lexique spécifique ou en identifient les bonnes définitions parmi plusieurs propositions, de manière encore restreinte mais nettement plus étendue que le groupe 1. Leurs connaissances apparaissent en effet plus développées en enseignement moral et

civique. Ils ont des acquis parcellaires sur des enseignements antérieurs à la troisième. La réussite de ces élèves s'exerce moins nettement dans le champ procédural (des capacités ou tâches) que sur des connaissances plus spécifiquement. Ils savent cependant exercer leur compréhension ou mobiliser des connaissances sur tous les supports documentaires. Ils commencent à traiter voire à interpréter (très rarement) les informations prélevées de ce matériau documentaire.

Groupe 3 (29,9 % des élèves)

Ce groupe maîtrise l'ensemble des compétences (capacités ou tâches et connaissances) fondamentales en histoire-géographie et en enseignement moral et civique. Ces élèves font preuve d'une bien meilleure maîtrise notionnelle et lexicale que les groupes de plus faible niveau. Ils savent identifier les notions usuelles et commencent à les caractériser. Leurs connaissances s'exercent non seulement sur le programme de troisième mais aussi sur certains contenus relevant d'enseignements suivis depuis la classe de cinquième au moins. Cependant leur maîtrise des repères historiques et géographiques est encore incomplète. Ils parviennent non seulement à identifier des informations, explicites ou implicites, mais aussi à traiter ou interpréter le matériau prélevé. La maîtrise de la compétence "identifier" est en effet très bonne, étayée par les connaissances lexicales, notionnelles et thématiques de ces élèves. Par ailleurs, la lecture de textes longs ou de propositions de réponses longues et complexes n'est plus un frein à la réussite. Pour traiter l'information, les élèves de ce groupe gèrent des documents et des tâches complexes. Ils donnent du sens aux documents et parviennent à généraliser : intituler, dégager l'idée principale d'un texte... Ils perçoivent les différences de points de vue. Ils peuvent reconnaître un récit d'historien dont ils distinguent certaines caractéristiques (argumentation qui prend appui sur des sources, effort d'objectivité...).

Groupe 4 (17,0 % des élèves)

Grâce à leurs connaissances approfondies, y compris sur les programmes des classes antérieures à la troisième, les élèves du groupe 4 font preuve d'une très large maîtrise des compétences évaluées, au sein desquelles l'équilibre devient vraiment probant entre tâches ou capacités et connaissances. Ils identifient ou définissent le vocabulaire spécifique et notionnel plus complexe en étant capables d'en appréhender et d'en distinguer largement les composantes et les implications. Ils font preuve d'une très large maîtrise des repères historiques du programme de troisième. Les élèves de ce groupe synthétisent des documents variés, longs et complexes... Ils donnent du sens au document ou à la situation considérée et ils en comprennent la portée en repérant souvent ce qui relève de l'implicite. Ils confrontent ces documents entre eux. Ils peuvent justifier, argumenter ou critiquer des affirmations ou un point de vue. Ces élèves commencent à appréhender la démarche des disciplines et à l'intégrer dans leur analyse. Ils

passent d'un langage à l'autre.

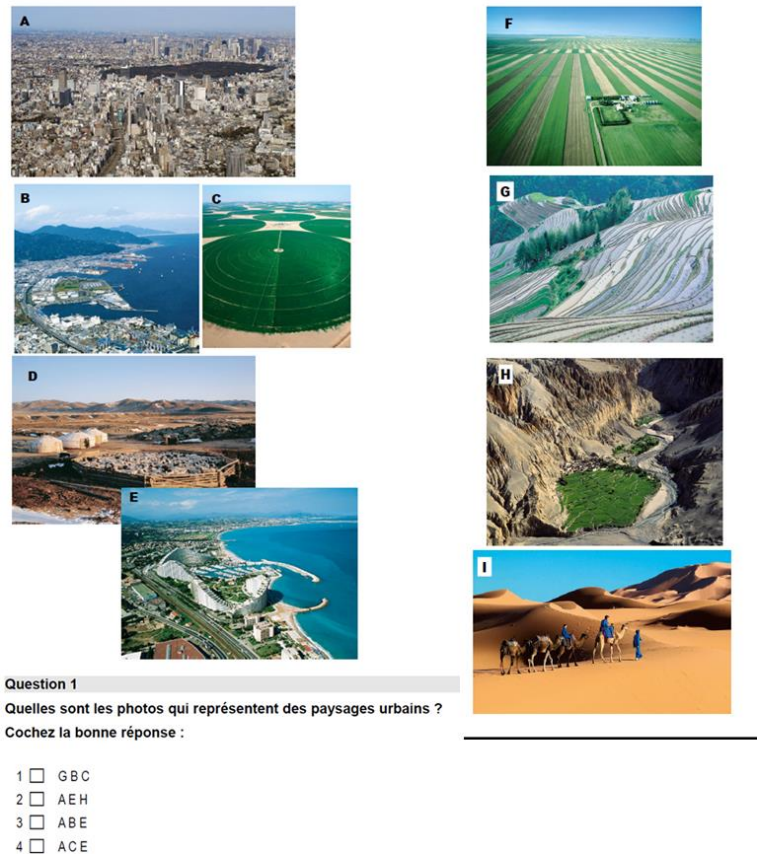
Groupe 5 (7,2 % des élèves)

Les élèves du groupe 5 ont acquis de manière très approfondie l'ensemble des connaissances et des compétences construites par l'enseignement de l'histoire-géographie et une culture civique déjà solide pendant leurs quatre années de scolarité au collège. Ces élèves vont directement au sens, à l'analyse critique, y compris à partir de leur lecture de textes complexes. De plus, ils répondent plutôt bien aux questions ouvertes : le travail de rédaction est plus abouti en histoire, de même que les tâches cartographiques.

5.3 Exemples d'items

5.3.1 Item caractéristique des groupes < 1 et 1

Figure 7 – Exemple groupe < à 1



Cet item est emblématique de ce que les groupes les plus faibles ont pu réussir. Il s'agit d'un item d'ancrage, autrement dit soumis à l'identique aux élèves interrogés en 2012 et à nouveau en 2017. C'est un des items les plus réussis par l'ensemble des élèves de tous les groupes, question pour laquelle le taux de réussite atteint 91 % en 2017 (résultat qu'on peut dire stable, puisque il était de 90 % en 2012).

L'item est positionné sur l'échelle au niveau groupe < 1 car la probabilité de réussite à cette question pour les élèves de ce groupe dépasse 50 % (elle est précisément de 55 %). Cela signifie que l'item est encore mieux réussi bien sûr

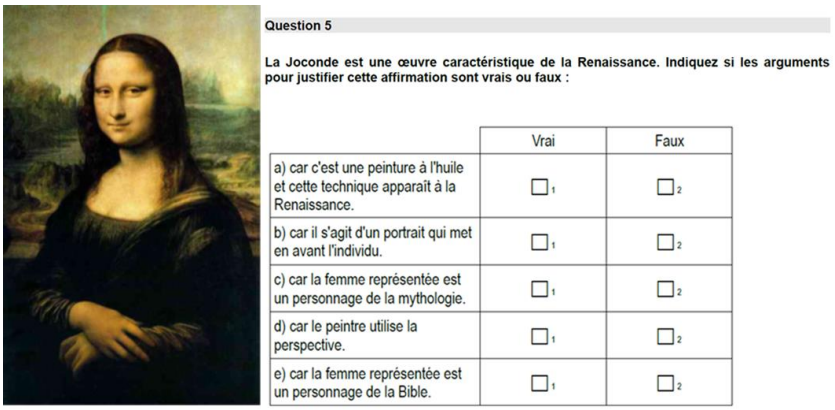
par les groupes de niveaux supérieurs sur l'échelle de performances CEDRE, à savoir le groupe 1 (probabilité de réussite de 62 %), le groupe 2 (83 %), le groupe 3 (94 %), le groupe 4 (98 %) et le groupe 5 (99 %).

On note que l'item relève de ce qu'on peut qualifier de reconnaissance. Il minimise la charge cognitive pour les élèves interrogés, en leur demandant simplement d'identifier des paysages urbains parmi ces propositions.

En ce qui concerne ces deux groupes les moins performants sur l'ensemble du test, seules quelques réponses de ce type, ponctuelles et dispersées, sont largement réussies. Elles concernent essentiellement des éléments de connaissances, certes très fragmentaires.

5.3.2 Item caractéristique du groupe 2

Figure 8 – Exemple groupe 2



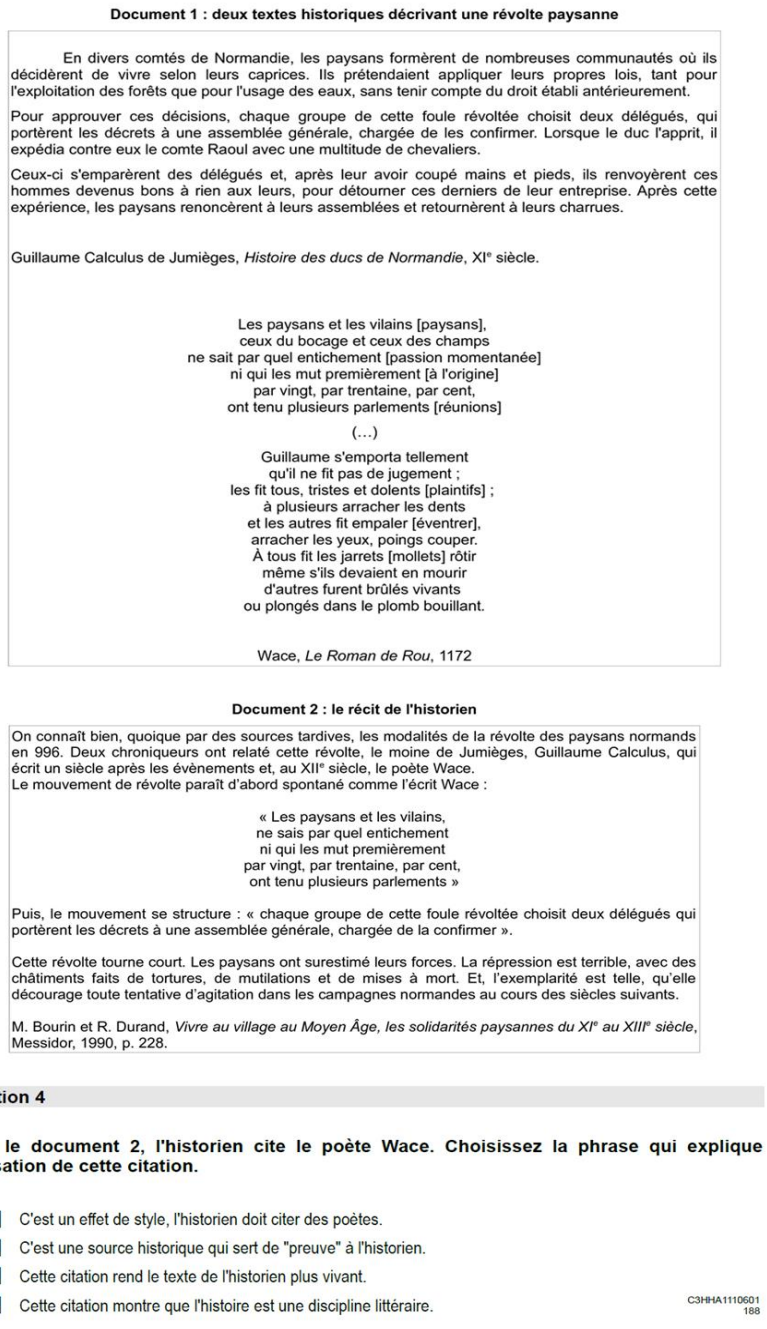
Cet item est assez emblématique des acquis des élèves du groupe 2, qui commencent à faire preuve d'une réussite bien plus large que les groupes < 1 et 1. Il s'agit d'un item d'ancrage (déjà proposé en 2012), pour lequel le taux de réussite global est de 76 % environ (en progrès de 5 points de pourcentage par rapport à 2012). La probabilité de réussite à cet item précis pour le groupe 2 atteint 54 %. Il s'agit ici d'un item discriminant car la probabilité de le réussir marque de véritables seuils d'un groupe à l'autre : 30 % pour le groupe 1 et 76 % pour le groupe 3, 90 % pour le groupe 4 et 96 % pour le groupe 5.

L'item interroge en série/tableau de type " vrai " ou " faux " les connaissances des élèves relatives à La Joconde en tant qu'oeuvre emblématique de la Renaissance. On peut reconnaître à cet item déjà une certaine difficulté, car ses propositions portent à fois sur La Joconde et sur la peinture de la Renaissance,

donc sur un premier niveau de mise en contexte. La proposition pour laquelle le taux de réussite est le moins élevé, de l'ordre de 73 % néanmoins, concerne la " mise en avant de l'individu " au travers de La Joconde et par la peinture de la Renaissance plus généralement. La réussite des élèves du groupe 2 sur l'ensemble du test, à l'image de ce qu'il en ressort pour cet item, autorise à parler d'un début de culture en histoire, en géographie et en enseignement moral et civique. Ces premiers jalons d'une culture portent cependant sur des savoirs encore peu élaborés pour la plupart. Les acquis du groupe 2 sont encore fragiles en effet, ils portent plus nettement sur des connaissances (dimension déclarative) que sur des tâches ou capacités (dimension procédurale). Les connaissances de ces élèves sont très largement restreintes au programme de troisième, même si on note quelques acquis relatifs aux enseignements des classes antérieures (ici, la Renaissance, étudiée en histoire en classe de cinquième). Il faut préciser aussi que les thématiques les plus largement réussies par ce groupe relèvent souvent de l'enseignement moral et civique. Il s'agit en particulier des valeurs et principes républicains, parmi lesquels la laïcité a fait l'objet d'un certain nombre d'items. En 2012 déjà, l'éducation civique avait été l'objet de résultats relativement stables, dans un contexte de baisse globale et marquée du score moyen des élèves sur l'ensemble du test par rapport à 2006.

5.3.3 Item caractéristique du groupe 3

Figure 9 – Exemple groupe 3



Document 1 : deux textes historiques décrivant une révolte paysanne

En divers comtés de Normandie, les paysans formèrent de nombreuses communautés où ils décidèrent de vivre selon leurs caprices. Ils prétendaient appliquer leurs propres lois, tant pour l'exploitation des forêts que pour l'usage des eaux, sans tenir compte du droit établi antérieurement.

Pour approuver ces décisions, chaque groupe de cette foule révoltée choisit deux délégués, qui portèrent les décrets à une assemblée générale, chargée de les confirmer. Lorsque le duc l'apprit, il expédia contre eux le comte Raoul avec une multitude de chevaliers.

Ceux-ci s'emparèrent des délégués et, après leur avoir coupé mains et pieds, ils renvoyèrent ces hommes devenus bons à rien aux leurs, pour détourner ces derniers de leur entreprise. Après cette expérience, les paysans renoncèrent à leurs assemblées et retournèrent à leurs charrues.

Guillaume Calculus de Jumièges, *Histoire des ducs de Normandie*, XI^e siècle.

Les paysans et les vilains [paysans],
ceux du bocage et ceux des champs
ne sait par quel entichement [passion momentanée]
ni qui les mut premièrement [à l'origine]
par vingt, par trentaine, par cent,
ont tenu plusieurs parlements [réunions]

(…)

Guillaume s'emporta tellement
qu'il ne fit pas de jugement ;
les fit tous, tristes et dolents [plaintifs] ;
à plusieurs arracher les dents
et les autres fit empaler [éventrer],
arracher les yeux, poings couper.
À tous fit les jarrets [mollets] rôtir
même s'ils devaient en mourir
d'autres furent brûlés vivants
ou plongés dans le plomb bouillant.

Wace, *Le Roman de Rou*, 1172

Document 2 : le récit de l'historien

On connaît bien, quoique par des sources tardives, les modalités de la révolte des paysans normands en 996. Deux chroniqueurs ont relaté cette révolte, le moine de Jumièges, Guillaume Calculus, qui écrit un siècle après les évènements et, au XII^e siècle, le poète Wace.
Le mouvement de révolte paraît d'abord spontané comme l'écrit Wace :

« Les paysans et les vilains,
ne sais par quel entichement
ni qui les mut premièrement
par vingt, par trentaine, par cent,
ont tenu plusieurs parlements »

Puis, le mouvement se structure : « chaque groupe de cette foule révoltée choisit deux délégués qui portèrent les décrets à une assemblée générale, chargée de la confirmer ».

Cette révolte tourne court. Les paysans ont surestimé leurs forces. La répression est terrible, avec des châtiments faits de tortures, de mutilations et de mises à mort. Et, l'exemplarité est telle, qu'elle décourage toute tentative d'agitation dans les campagnes normandes au cours des siècles suivants.

M. Bourin et R. Durand, *Vivre au village au Moyen Âge, les solidarités paysannes du XI^e au XIII^e siècle*, Messidor, 1990, p. 228.

Question 4

Dans le document 2, l'historien cite le poète Wace. Choisissez la phrase qui explique l'utilisation de cette citation.

1 ☐ C'est un effet de style, l'historien doit citer des poètes.
2 ☐ C'est une source historique qui sert de "preuve" à l'historien.
3 ☐ Cette citation rend le texte de l'historien plus vivant.
4 ☐ Cette citation montre que l'histoire est une discipline littéraire.

C3HHA1110601
188

Cet exemple illustre le fait qu'à partir du groupe 3, la lecture de textes longs, au vocabulaire difficile, ou de propositions longues ou complexes, n'est plus autant un frein à la réussite. L'item relève d'une unité (ensemble de quelques items successifs portant sur une thématique) conçue pour l'évaluation de 2012, puis à nouveau intégrée dans le corpus de l'évaluation de 2017 au titre de l'ancrage.

Les élèves y sont dans un premier temps confrontés à deux sources historiques (document 1). Ces deux textes du XI^{e} et du XII^{e} s. décrivent une révolte paysanne survenue en Normandie à la fin du X^{e} s. Les élèves sont ensuite interrogés à propos d'un récit d'historiens (document 2) établi à partir de ces deux sources médiévales. Ces documents, qui renvoient au programme de la classe de cinquième n'ont, selon toute vraisemblance, que très rarement été étudiés en classe.

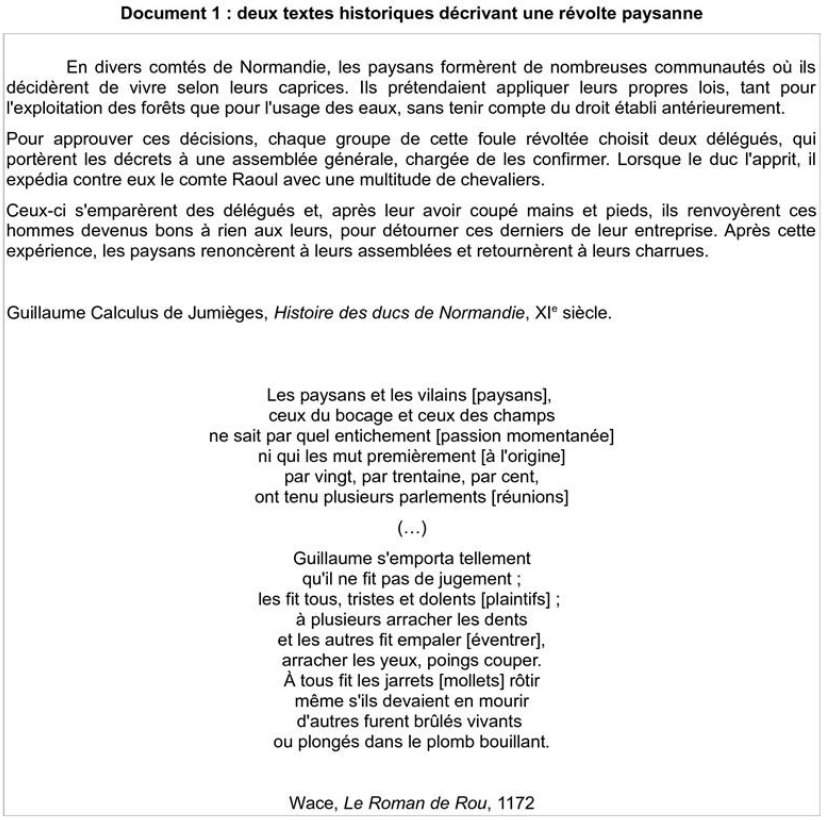
Après quelques questions relevant du prélèvement d'informations pour évaluer la compréhension de ces textes par les élèves, d'autres items portent sur l'interprétation et la critique du document. Puis une dernière question renvoie aux composantes du récit historique. Les items de cette unité se positionnent sur l'échelle de performances, selon le cas, au niveau du groupe 3 ou du groupe 4...

Cette question 4 met l'accent sur l'utilisation de la source et de la citation comme " preuve " soutenant la thèse de l'historien, qui se distingue des usages plus littéraires. Cet item a été réussi par 56 % des élèves (progrès de l'ordre de 4 points par rapport à 2012). On note que la dernière proposition (" ... montre que l'histoire est une discipline littéraire "), qu'on peut tenir a priori pour un distracteur attrayant, n'a été choisie que par 10 % environ des répondants en 2012 comme en 2017, derrière la troisième proposition, cochée par 16 % des élèves (" ... rend le texte de l'historien plus vivant ") et à peine plus que la première proposition, sélectionnée par un peu plus de 8 % des collégiens (" l'historien doit citer des poètes ") alors qu'on aurait pu la croire bien moins attractive. On mesure pour cet exemple, au passage, quant à l'appréhension de la démarche historienne, une nette distance entre le savoir assimilé par les élèves au collège d'une part, et la réflexion savante et universitaire à ce sujet d'autre part, tant la question de la dimension littéraire de l'écriture historienne continue de faire débat entre les spécialistes (cf. notamment Ivan Jablonka, *L'histoire est une littérature contemporaine. Manifeste pour les sciences sociales*, Seuil, 2014, qui qualifie l'écriture de l'historien de " littérature référentielle "). Il reste, cependant, que l'usage de la citation par le texte que les élèves devaient ici considérer relève indiscutablement du régime de la preuve et ne prêtait guère à confusion pour des élèves en fin de collège quant au choix de la bonne réponse à apporter à cette question 4.

Il s'agit en outre d'un item discriminant, qualité qui atteste de sa pertinence à refléter ce qui est effectivement enseigné et appris. En d'autres termes, la question a été très largement réussie par les élèves les plus en réussite sur l'ensemble du test, bien moins par les élèves en difficulté, avec une pente croissante et cohérente de l'un à l'autre des groupes de niveaux, tout en marquant dans le cas présent un saut au niveau du groupe 3. La probabilité de réussite passe en effet de 21 % pour le groupe 2 à 53 % pour le groupe 3.

5.3.4 Item caractéristique du groupe 4

Figure 10 – Exemple groupe 4 - 1



Document 1 : deux textes historiques décrivant une révolte paysanne

En divers comtés de Normandie, les paysans formèrent de nombreuses communautés où ils décidèrent de vivre selon leurs caprices. Ils prétendaient appliquer leurs propres lois, tant pour l'exploitation des forêts que pour l'usage des eaux, sans tenir compte du droit établi antérieurement.

Pour approuver ces décisions, chaque groupe de cette foule révoltée choisit deux délégués, qui portèrent les décrets à une assemblée générale, chargée de les confirmer. Lorsque le duc l'apprit, il expédia contre eux le comte Raoul avec une multitude de chevaliers.

Ceux-ci s'emparèrent des délégués et, après leur avoir coupé mains et pieds, ils renvoyèrent ces hommes devenus bons à rien aux leurs, pour détourner ces derniers de leur entreprise. Après cette expérience, les paysans renoncèrent à leurs assemblées et retournèrent à leurs charrues.

Guillaume Calculus de Jumièges, *Histoire des ducs de Normandie*, XIe siècle.

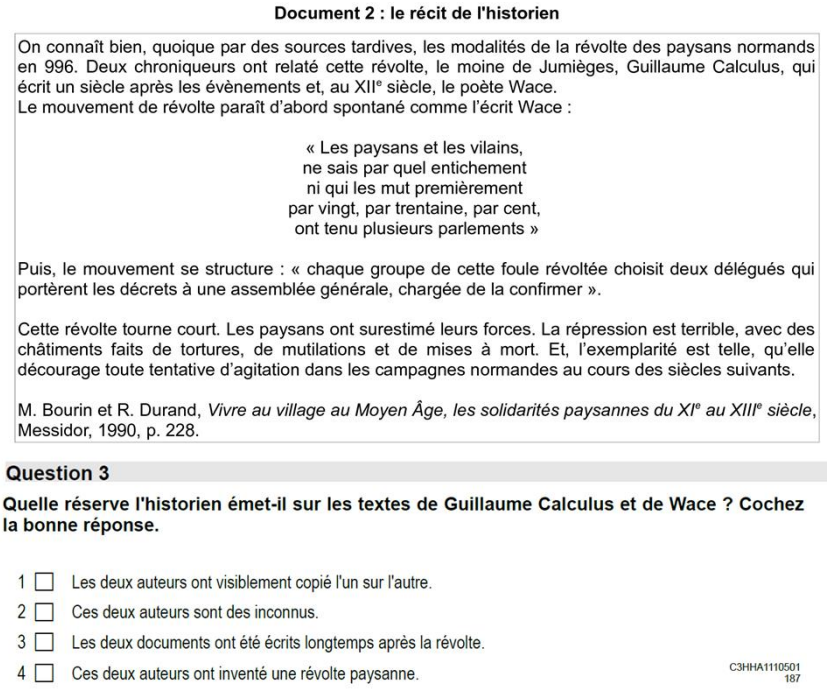
Les paysans et les vilains [paysans],
ceux du bocage et ceux des champs
ne sait par quel entichement [passion momentanée]
ni qui les mut premièrement [à l'origine]
par vingt, par trentaine, par cent,
ont tenu plusieurs parlements [réunions]

(...)

Guillaume s'emporta tellement
qu'il ne fit pas de jugement ;
les fit tous, tristes et dolents [plaintifs] ;
à plusieurs arracher les dents
et les autres fit empaler [éventrer],
arracher les yeux, poings couper.
À tous fit les jarrets [mollets] rôtir
même s'ils devaient en mourir
d'autres furent brûlés vivants
ou plongés dans le plomb bouillant.

Wace, *Le Roman de Rou*, 1172

Figure 11 – Exemple groupe 4 - 2



Document 2 : le récit de l'historien

On connaît bien, quoique par des sources tardives, les modalités de la révolte des paysans normands en 996. Deux chroniqueurs ont relaté cette révolte, le moine de Jumièges, Guillaume Calculus, qui écrit un siècle après les évènements et, au XII[e] siècle, le poète Wace.
Le mouvement de révolte paraît d'abord spontané comme l'écrit Wace :

« Les paysans et les vilains,
ne sais par quel entichement
ni qui les mut premièrement
par vingt, par trentaine, par cent,
ont tenu plusieurs parlements »

Puis, le mouvement se structure : « chaque groupe de cette foule révoltée choisit deux délégués qui portèrent les décrets à une assemblée générale, chargée de la confirmer ».

Cette révolte tourne court. Les paysans ont surestimé leurs forces. La répression est terrible, avec des châtiments faits de tortures, de mutilations et de mises à mort. Et, l'exemplarité est telle, qu'elle décourage toute tentative d'agitation dans les campagnes normandes au cours des siècles suivants.

M. Bourin et R. Durand, *Vivre au village au Moyen Âge, les solidarités paysannes du XI[e] au XIII[e] siècle*, Messidor, 1990, p. 228.

Question 3

Quelle réserve l'historien émet-il sur les textes de Guillaume Calculus et de Wace ? Cochez la bonne réponse.

1 ☐ Les deux auteurs ont visiblement copié l'un sur l'autre.
2 ☐ Ces deux auteurs sont des inconnus.
3 ☐ Les deux documents ont été écrits longtemps après la révolte.
4 ☐ Ces deux auteurs ont inventé une révolte paysanne.

C3HHA1110501
187

Cet item est tiré de la même unité que l'exemple précédent, significatif de la réussite du groupe 3. Ici, la question 3, emblématique de la réussite du groupe 4, concerne la compréhension et l'interprétation du document en histoire. Elle vérifie la capacité des élèves à comprendre une critique, ou plus exactement une "réserve", émise par l'historien à l'égard de ses sources (les deux textes réunis en document 1 et sur lesquels les élèves ont par ailleurs été interrogés).

Au début de leur texte en effet, les historiens pointent que leurs sources sont " tardives " au sujet cette révolte survenue à la fin du X[e] s. en Normandie. La bonne réponse est donc la troisième proposition. Le résultat est stable par rapport à 2012.

Les première et dernière propositions, erronées, ont été choisies par plus d'un quart des élèves tous groupes confondus. On peut interroger ce tropisme des propositions 1 et 4. Il est certes difficile de faire la part entre les élèves qui n'ont tout simplement pas relevé l'information explicite dans le texte au sujet du caractère tardif des sources, et ceux qui l'ont perçue et comprise mais qui ont été conduits à surinterpréter cette "réserve" émise par les historiens. L'erreur a surtout porté sur la dernière proposition (14 % des réponses) qui disqualifie en

quelque sorte les deux sources en question (" ont inventé une révolte "). Or, les élèves du groupe 3, eux-mêmes encore en difficulté sur cet item (probabilité de réussite de 50 %), réussissent pourtant largement les autres items de prélèvement d'informations dans des textes. Autrement dit, bon nombre d'élèves, jusqu'au groupe 3 au moins, semblent évaluer la valeur de la source en histoire uniquement à l'aune de sa dimension contemporaine des faits relatés, ou au contraire de son caractère tardif... Ce constat ressort aussi de l'analyse des résultats pour d'autres items proches de cet exemple.

Quoiqu'il en soit, cet item révèle que les élèves du groupe 4, qui atteignent une probabilité de réussite de 65 % à cet item, répondent aux enjeux d'un esprit critique tels que les synthétisent Jérôme Grondeux (historien et IGEN) : apprendre à " ne pas précipiter son jugement ", car " en histoire, critiquer un document, ce n'est pas le disqualifier, c'est rechercher l'intérêt qu'il a et ce qu'il peut nous apprendre d'intéressant avec un certain degré de certitude " ... (" L'esprit critique ", sur Eduscol, page mise à jour le 8 décembre 2016, et " Former l'esprit critique des élèves ", sur Eduscol également, page mise à jour le 10 janvier 2018).

On perçoit ainsi qu'à partir de ce groupe 4, avec une aisance encore plus marquée pour le groupe 5 bien sûr, les élèves commencent à appréhender la démarche des disciplines et à l'intégrer dans leur analyse du document. Ils font plus largement preuve de leur capacité à synthétiser et à interpréter des documents, de toutes natures et complexes. Ils leur donnent du sens et ils commencent à comprendre leur portée, en repérant souvent ce qui relève de l'implicite. Ils mobilisent pour cela leur connaissance du contexte, ou ils mettent en perspective les cas et les exemples proposés en géographie, et ils confrontent ces documents entre eux. Ils peuvent également justifier, argumenter ou critiquer une affirmation ou un point de vue.

Il faut également noter que les élèves des groupes 4 et 5 se distinguent des groupes de niveaux plus faibles par leur capacité à passer d'un langage à l'autre, au moyen d'une tâche cartographique ou par une rédaction, de type récit historique.

5.3.5 Item caractéristique du groupe 5

Figure 12 – Exemple groupe 5 - 1

Figure 13 – Exemple groupe 5 - 2

	François Ier	Louis XIV	Périclès	Philippe Auguste
Monument historique A	☐ 1	☐ 2	☐ 3	☐ 4
Monument historique B	☐ 1	☐ 2	☐ 3	☐ 4
Monument historique C	☐ 1	☐ 2	☐ 3	☐ 4
Monument historique D	☐ 1	☐ 2	☐ 3	☐ 4

Des élèves du groupe 5 on peut dire qu'ils sont ceux qui maîtrisent les repères chronologiques et spatiaux liés au programme de 3e, habileté déjà marquée au niveau du groupe 4, et aux enseignements des classes antérieures, ce qui caractérise davantage et plus spécifiquement ce groupe 5 le plus en réussite sur l'ensemble du test.

Les élèves du groupe 5, plus particulièrement, parviennent à appréhender ces

repères à la fois sur le temps long et sur le temps plus court d'une période. Par exemple, ils peuvent situer dans le temps de grands personnages étudiés depuis la 6e, leur associer régimes politiques ou monuments correspondants, comme ici sur cet exemple. Par ailleurs, ces élèves ont des connaissances assez fines, en fin de troisième, sur la Révolution française étudiée en classe de quatrième.

6 Variables contextuelles et non cognitives

6.1 Variables sociodémographiques et indice de position sociale

Un certain nombre de variables sociodémographiques permettent d'enrichir l'analyse des résultats. Le score moyen des élèves est ainsi analysé en fonction du genre, du retard scolaire et quand les effectifs le permettent en fonction du secteur d'enseignement. Le lecteur est invité à consulter la Note d'Information pour plus de détails (Ninnin & Berton, 2018).

L'indice de position sociale mesure la proximité au système scolaire du milieu familial de l'enfant. Cet indice peut se substituer à la profession des parents pour mieux expliquer les parcours et la réussite scolaire de leurs enfants. Il consiste en une transformation des PCS en valeur numérique (Rocher, 2016).

Il n'a été possible d'établir des comparaisons qu'en termes de niveau social des écoles, et non au niveau individuel. En effet, en 2017, la PCS des parents est disponible pour chaque élève, mais elle ne l'était pas dans les cycles antérieurs. Pour chaque établissement des échantillons de 2006, 2012 et 2017, la moyenne de l'indice de position socio-scolaire a été calculée et la population a ensuite été découpée en quatre groupes selon les quartiles (tableau 17).

Tableau 17 – Score moyen selon l'indice de position sociale moyen de l'établissement (CEDRE 2017)

Indice moyen école	Année	Répartition (%)	Score moyen	Écart type
1er quart	2006	24.8	233	48
1er quart	2012	24.6	**224**	46
1er quart	2017	24.5	228	47
2e quart	2006	24.7	246	47
2e quart	2012	25.1	**236**	46
2e quart	2017	25.3	241	45
3e quart	2006	24.2	255	50
3e quart	2012	24.9	**241**	49
3e quart	2017	25.0	247	45
4e quart	2006	26.2	265	50
4e quart	2012	25.4	258	48
4e quart	2017	25.1	264	49

Note de lecture : en 2012, le score moyen des élèves appartenant au quart des collèges les plus défavorisées (1er quart) diminue de 9 points par rapport à 2006. Les évolutions significatives sont indiquées en gras.

6.2 Élaboration des questionnaires de contexte

Pour pouvoir davantage enrichir l'analyse des résultats, deux questionnaires de contexte ont été élaborés. Un questionnaire élève a été ajouté à la fin du cahier d'évaluation et un questionnaire enseignant était adressé aux enseignants des classes participantes à l'évaluation. Ces questionnaires ont été élaborés en collaboration avec des chercheurs et des spécialistes en sciences de l'éducation.

Le questionnaire enseignant interroge les enseignants sur leur niveau de formation et leur ancienneté. Ce questionnaire inclut aussi des questions sur les pratiques pédagogiques, les stratégies d'enseignement, le sentiment d'efficacité personnelle etc.

Le questionnaire élève interroge des dimensions dites conatives intéressantes à mettre en lien avec le score obtenu à l'épreuve - temps de travail personnel estimé par l'élève, type de travail personnel le plus demandé, pratiques culturelles en lien avec les disciplines évaluées... De plus, les élèves sont demandés d'évaluer la difficulté de l'épreuve et leur degré d'implication à faire le test.

6.3 Motivation des élèves face à la situation d'évaluation

Les évaluations standardisées des élèves, telles que CEDRE ou PISA, renvoient à des enjeux politiques croissants, alors qu'elles restent à faible enjeu pour les élèves participants. Dans le système éducatif français, où la notation tient une place prépondérante, la question de la motivation des élèves face à ces évaluations mérite d'être posée.

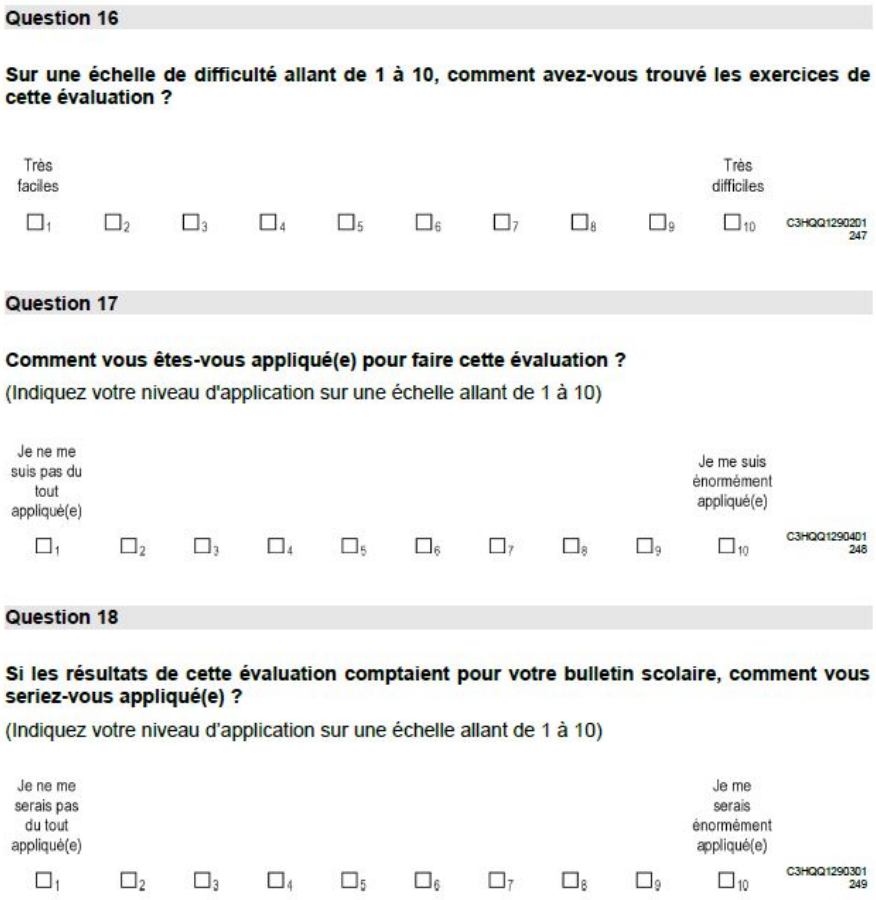
Un instrument pour mesurer la motivation a été adapté à partir du « thermomètre d'effort » proposé dans PISA (Keskpaik. & Rocher, 2015). Cet instrument (cf. figure 14) a été introduit dans plusieurs évaluations conduites au niveau national par la DEPP, y compris dans CEDRE maîtrise de la langue. Les données recueillies permettent de distinguer la motivation de l'élève de la difficulté perçue du test, et ainsi de mieux appréhender le lien entre la motivation des élèves français et leur performance. L'analyse de ces données renseigne en outre sur le rôle de certaines caractéristiques, des élèves ou des évaluations elles-mêmes, dans le degré de motivation à répondre aux questions de l'évaluation.

Le tableau 18 présente les grands résultats de cet instrument.

Tableau 18 – Résultats de l'instrument de mesure de la motivation au test (CEDRE 2017)

	Moyenne	Erreur standard
Comment avez-vous trouvé les exercices de cette évaluation ?	5,2	1,99
Comment vous êtes-vous appliqué(e) pour faire cette évaluation ?	6,3	2,23
Si les résultats de cette évaluation comptaient pour votre bulletin scolaire, comment vous seriez-vous appliqué(e) ?	8,9	1,61

Figure 14 – Instrument de mesure de la motivation au test



Question 16

Sur une échelle de difficulté allant de 1 à 10, comment avez-vous trouvé les exercices de cette évaluation ?

Très faciles ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9 ☐10 Très difficiles

C3HQQ1290201
247

Question 17

Comment vous êtes-vous appliqué(e) pour faire cette évaluation ?

(Indiquez votre niveau d'application sur une échelle allant de 1 à 10)

Je ne me suis pas du tout appliqué(e) ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9 ☐10 Je me suis énormément appliqué(e)

C3HQQ1290401
248

Question 18

Si les résultats de cette évaluation comptaient pour votre bulletin scolaire, comment vous seriez-vous appliqué(e) ?

(Indiquez votre niveau d'application sur une échelle allant de 1 à 10)

Je ne me serais pas du tout appliqué(e) ☐1 ☐2 ☐3 ☐4 ☐5 ☐6 ☐7 ☐8 ☐9 ☐10 Je me serais énormément appliqué(e)

C3HQQ1290301
249

7 Annexe

Certification AFNOR pour les évaluations CEDRE

La DEPP est engagée dans un processus de certification. Elle a obtenu en mars 2015 la certification pour les évaluations CEDRE.

Les finalités de la certification

Les finalités sont les suivantes :

- inscrire les processus d'évaluation dans une dynamique pérenne d'amélioration continue ;
- renforcer la prise en compte des attentes des usagers dans la formalisation des objectifs des évaluations et la restitution de leurs résultats ;
- faire reconnaître par une certification de service la qualité du service rendu et la continuité du respect des engagements pris.

Les enjeux pour la DEPP

Il y a deux enjeux forts pour la DEPP, l'un interne, l'autre externe :

- améliorer les processus de construction des instruments d'évaluation des acquis des élèves, fiabiliser ces processus par une démarche de contrôle-qualité ;
- valoriser l'enquête CEDRE comme un standard de qualité procédurale dans le domaine de l'évaluation.

Plus spécifiquement, le projet de certification des évaluations CEDRE est porteur d'enjeux pour la DEPP en termes de communication sur la validité scientifique, la sincérité, l'objectivité et la fiabilité des évaluations, ainsi que sur l'éthique et le professionnalisme des équipes.

La démarche qualité

Elle est fondée sur un référentiel élaboré sur mesure, selon une démarche officielle reconnue par les services publics et en lien avec les représentants des utilisateurs du service et les professionnels. La transparence vis-à-vis des usagers est assurée par la communication des résultats des enquêtes de satisfaction annuelles.

Les engagements de service

Le référentiel d'engagements comporte 18 engagements (cf. encadré page suivante).

Les engagements de service de la DEPP

Des objectifs clairs et partagés

Nous associons les parties intéressées à la définition de notre programme d'évaluation.

Nous formalisons dans un " cadre d'évaluation " les résultats attendus et les paramètres techniques de l'évaluation, ses délais et les limites associées aux moyens mis en œuvre.

Des évaluations fondées sur l'expertise pédagogique

Nous définissons avec les parties intéressées les acquis à évaluer et les mesurons en intégralité.

Nous mobilisons, tout au long de l'évaluation, un groupe expérimenté composé d'enseignants de terrain, de formateurs, d'inspecteurs et de chercheurs.

Tous nos items sont testés, analysés et validés avec le groupe expert avant d'être utilisés dans le cadre d'une évaluation.

Les meilleures pratiques méthodologiques et statistiques au service de l'objectivité

Afin de garantir l'application des meilleures méthodes statistiques, nous prenons en compte avec exigence les principes du " Code de bonnes pratiques de la statistique européenne ".

Nous tirons un échantillon représentatif garantissant le maximum de précision de mesure, à partir du plan de sondage défini dans le respect du " cadre d'évaluation ".

Nous garantissons l'objectivité et la qualité des données recueillies par la standardisation des processus d'administration et de correction des tests.

Une mesure fiable et des comparaisons temporelles pertinentes

Afin de garantir l'application des meilleures méthodes psychométriques, nous prenons en compte avec exigence les recommandations internationales sur l'utilisation des tests.

Nous analysons les réponses apportées par les élèves aux items afin d'en garantir la validité psychométrique.

Nous modélisons une échelle de compétences servant de référence et offrons des comparaisons temporelles fiables et lisibles.

Nous caractérisons les niveaux de cette échelle et déterminons avec le groupe expert les seuils de maîtrise des compétences évaluées, permettant de vous décrire en détail les performances des élèves.

Des analyses enrichies par des données de contexte

Nous systématisons le recueil d'informations standardisées relatives aux élèves et à leur environnement scolaire et social, dans le respect le plus strict des règles de confidentialité.

Nous éclairons les résultats de nos évaluations par la mise en relation des scores avec ces données.

Transparence des méthodes et partage des résultats

Nous publions et présentons les résultats de chacune de nos évaluations.

Nous mettons à disposition un rapport technique précisant les méthodes utilisées dans le cadre de l'évaluation.

Nous participons, dans le cadre de conventions collaboratives, à des analyses complémentaires des données que nous produisons.

Références

Ardilly, P. (2006). *Les techniques de sondage.* Technip.

Christine, M., & Rocher, T. (2012, janvier). Construction d'échantillons astreints à des conditions de recouvrement par rapport à un échantillon antérieur et à des conditions d'équilibrage par rapport à des variables courantes : aspects théoriques et mise en œuvre dans le cadre du renouvellement des échantillons des enquêtes d'évaluation des élèves. In *Journées de méthodologie statistique.* Paris.

Garcia, E., Le Cam, M., & Rocher, T. (2015). Méthodes de sondage utilisées dans les programmes d'évaluation des élèves. *Éducation et Formations*, *85-86*, 101-117.

Keskpaik., S., & Rocher, T. (2015). La motivation des élèves français face à des évaluations à faibles enjeux. comment la mesurer ? son impact sur les réponses. *Education et formations*, *85-86*, 119-139.

Ninnin, L.-M., & Berton, S. (2018). CEDRE 2006-2012-2017, histoire, géographie, enseignement moral et civique en fin de collège : un progrès global des acquis des élèves après une baisse constatée en 2012. *Note d'information*, *16*.

Rocher, T. (1999). *Psychométrie et théorie des sondages* (Mémoire de Master non publié). Université Paris VI.

Rocher, T. (2013). *Mesure des compétences : les méthodes se valent-elles ? questions de psychométrie dans le cadre de l'évaluation de la compréhension de l'écrit* (Thèse de doctorat non publiée). Université Paris-Ouest.

Rocher, T. (2015). Mesure des compétences : méthodes psychométriques utlisées dans le cadre des évaluations des élèves. *Éducation et Formations*, *86-87*, 37-60.

Rocher, T. (2016). Construction d'un indice de position sociale des élèves. *Éducation et Formations*, *90*, 5-27.

Rousseau, S., & Tardieu, F. (2004). *La macro sas cube d'échantillonnage équilibré. documentation de l'utilisateur.* Paris : INSEE.

Sautory, O. (1993). La macro calmar. redressement d'un échantillon par calage sur marges. *Série des documents de travail de l'INSEE*, *Document F9310*.

Smith, R., Schumaker, R., & Bush, J. (1998). Using item mean squares to evaluate fit to the rasch model. *Journal of Outcome Measurement*, *2 n° 1*, 66-78.

Tillé, Y. (2001). *Théorie des sondages. echantillonnage et estimation en populations finies. cours et exercices avec solution.* Paris : Dunod.

Trosseille, B., & Rocher, T. (2015). Les évaluations standardisées des élèves. perspective historique. *Éducation et Formations*, *85-86*, 15-35.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54 n°3*, 427-450.

Liste des tableaux

Table des figures